Regression Analysis on Total Annual Mackerel and Horse Mackerel Egg Production

Jiamu Zhao*

Faculty of Mathematics, University of Waterloo, Waterloo, Canada

*j449zhao@uwaterloo.ca

Abstract. The International Council for the Exploration of the Sea (ICES) plays a crucial role in marine science, focusing on the sustainable management and conservation of marine biodiversity. This paper presents a detailed analysis of data from the Mackerel and Horse Mackerel Egg Surveys (WGMEGS), conducted under the ICES. Specifically, we explore the 2010 datasets, which estimate the spawning-stock biomass of mackerel and horse mackerel in the Northeast Atlantic and North Sea. Our research is structured around three primary objectives: (1) Estimating the total annual egg production for mackerel and horse mackerel using statistical models, (2) Identifying key features among selected covariates that significantly impact egg production, and (3) Analyzing the differences between two surveys conducted in the same year. The outcomes of our study aim to provide insights that are critical for the effective management and conservation of these vital fish species, contributing to the broader goal of safeguarding marine biodiversity.

Keywords: Modeling and simulation, Model development and analysis, Model verification and validation

1 Introduction

1.1 Dataset Background

The International Council for the Exploration of the Sea (ICES) serves as a pivotal intergovernmental organization in marine science, dedicated to fulfilling societal needs by providing unbiased evidence regarding the health and sustainable management of our seas and oceans. ICES conducts various studies, including the "Mackerel and Horse Mackerel Egg Surveys," overseen by the Working Group on Mackerel and Horse Mackerel Egg Surveys (WGMEGS). This survey, initiated in 1968, focuses on the Northeast Atlantic and the North Sea, and it is conducted triennially. The aim is to safeguard marine biodiversity by estimating the spawning-stock biomass of both mackerel and horse mackerel as shown in an online reference [4]. The datasets we selected originates from mackerel egg and horse mackerel surveys conducted in 2010. We will specifically conduct our regression analysis on that dataset and study the following research questions.

A. K. Draman Mud et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Social Sciences (ICBDSS 2024)*, Advances in Computer Science Research 116,
https://doi.org/10.2991/978-94-6463-562-1_35

1.2 Proposed Research Question

We have three proposed research questions in total.

1.2.1 Estimation on Egg Production.

We aim to calculate the total annual egg production for mackerel and horse mackerel using statistical models, utilizing data gathered from both surveys. This approach will allow for a precise estimation of reproductive output, which is critical for the effective management and conservation of these species.

1.2.2 Feature Importance.

We will detect any important features among the selected covariates.

1.2.3 Difference in Two Surveys.

We will determine if there is any difference between horse mackerel egg production and mackerel egg production.

2 Data Description

2.1 Source of Data

The dataset used in this study is available through the "Generalized Additive Models datasets" by installing the R package "gamair" as shown in the dataset [1]. Within this package, the data can be accessed using "data(meh)" and "data(med)", which we utilized in our project. Alternatively, the International Council for the Exploration of the Sea (ICES) provides an open-access resource on their website, where all survey-related data is freely available to the public.

2.2 Variables and Sample Size

The number of stage I eggs is our target variable, referred to as "count." Since the dataset consists only of integer values, we apply a logarithmic transformation to make the data continuous. The remaining variables are explanatory and include the position of the sample station, temperature measurements, salinity, seabed depth, the volume of sampled water, the responsible country, the type of sampling gear used, the ID of ships and samples, and the time of sample collection. We combine the two datasets and create a new indicator variable called "Type," denoting the species of the egg collection.

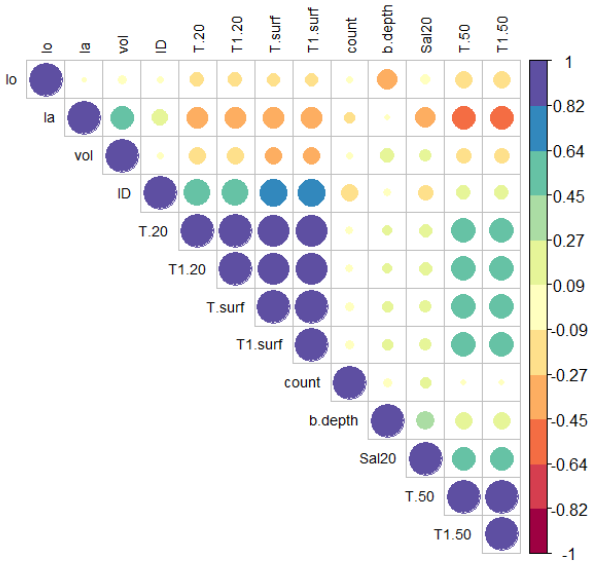

Fig. 1. Correlation plot.

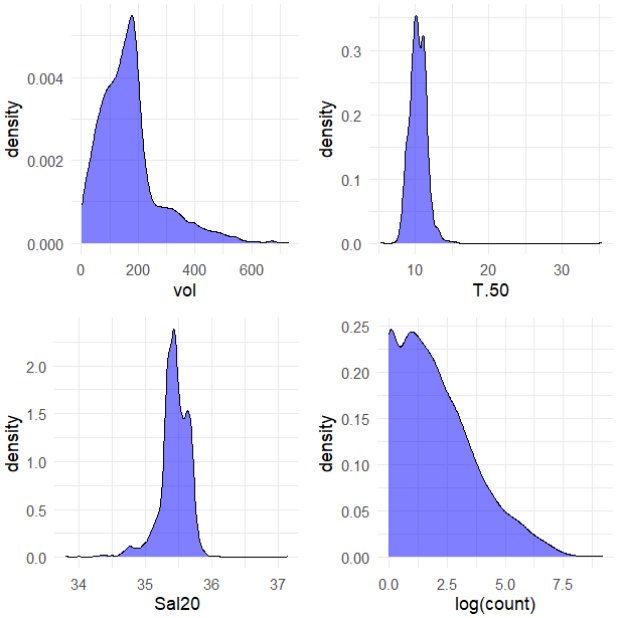

Fig. 2. Density plots.

3 Methodology

3.1 Data Preprocessing

We mainly use R as our tool for analyzing our dataset [5]. We began our exploratory data analysis with a correlation plot. As shown in Figure 1, some explanatory variables are highly correlated, particularly the temperature variables. Additionally, "b.depth" does not correlate with any other variables. Consequently, we removed these variables, except for two temperature variables, before training any statistical models. Therefore, the final variable selection includes "la," "lo," "vol," "Sal," "T.surf," "T.50," "country," "gear," and "Type."

As we can see from Figure 2, the distribution of the "vol" variable begins around 0 and rises sharply, peaking between 50 and 100. After this peak, the density gradually decreases, extending towards higher values but never reaching 0, resulting in a long right tail. The "T.50" variable's distribution is sharply peaked with a very narrow spread, having the highest density around the value of 10. The "Sal20" variable's distribution starts near 34, and then declines, indicating that most values are clustered around the peak. For the "log(count)" variable, the distribution starts near 0, rises quickly to a peak between 1 and 2, and then declines steadily.

Since the explanatory variables contain both numerical variables and categorical variables, we apply one-hot encoding on categorical variables for transforming them into numerical values. Though the calculation of training set becomes intensive, we use such algorithm for removing the ordinality relationship between each categorical values.

3.2 Planned Model

Given our research question, we plan to fit the following regression models on the selected dataset.

3.2.1 LASSO Regression.

$$\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j\,x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

As usual, we assume a linear relationship between the predictors and the response variable in our dataset. Thus, we fit a LASSO regression, which includes both the residual sum of squares and the penalty term by a technical report [2]. By doing so, LASSO regression achieves variable selection and regularization, improving model interpretability and potentially reducing overfitting.

The table below displays the non-zero coefficients under LASSO regression. For our training set, we applied 10-fold cross-validation on LASSO regression to find the lambda value that minimizes the mean squared error. Consequently, we selected $\lambda = 0.00729$ for our LASSO regression model.

As illustrated in Table 1, the type of fish has a strong effect on the dependent variable, with -1.130 indicating that the egg production of horse mackerel is greater than that of mackerel. "Sal20" and some countries also have relatively large estimates, suggesting their contributions to the model are substantial enough to withstand penalization.

Table 1. LASSO regression non-zero coefficients.

Intercept	la	lo	vol	Sal20
-23.89	-0.05204	0.002578	0.001389	0.7862
Type	T.surf	T.50	countryICELAND	countryIRELAND
-1.130	-0.01697	0.03579	-0.05216	0.09755
countryNORWAY	countrySPAIN	gearBgn	gearBGN60	gearGulf7
-0.1943	-0.6427	-0.2271	-0.0007583	0.2402

3.2.2 Tree.

A decision tree is a model which assigns discrete predictions based upon a partition of the sample space. Due to the fact that decision trees are easy to interpret, visualize, and simple to understand and implement, we fit a tree model on our dataset.

My fitted decision tree plot uses tree plot method, given from a technical report [7], is shown in Figure 3. We used our training set as the dataset for the tree model with the default values of maximum depth and minimum samples per leaf.

As we can see from the plot, "T.50", "vol", "Sal20", "T.surf", and "la" are relatively important features in our tree model. Specifically, most observations fall into the first and fifth leaf nodes if counted from left to right: 631 observations are classified by "T.50 < 11," and 450 observations are classified by "Sal20 < 35."

3.2.3 Random Forest.

A random forest is an ensemble of multiple decision trees, where each tree is built on a subset of the data and a subset of the features by a technical report [3]. The final prediction is made by averaging the predictions of all individual trees. Since the random forest model generally provides better predictive accuracy due to the ensemble approach, we fitted our random forest model on the dataset to see if it performs better than our tree model.

In our random forest model, we set "ntree" to 100 (default value of "ntree)" and applied grid search on the "mtry" parameter to find the value that minimizes RMSE. As shown in the Table 2 below, RMSE first decreases as "mtry" increases, and then increases when "mtry" have reached a certain threshold. We calculate both RMSE based on training set and test set, and select the "mtry" with the lowest RMSE value. So we set "mtry" to be 50 to fit our random forest model, though overfitting exists in this model.

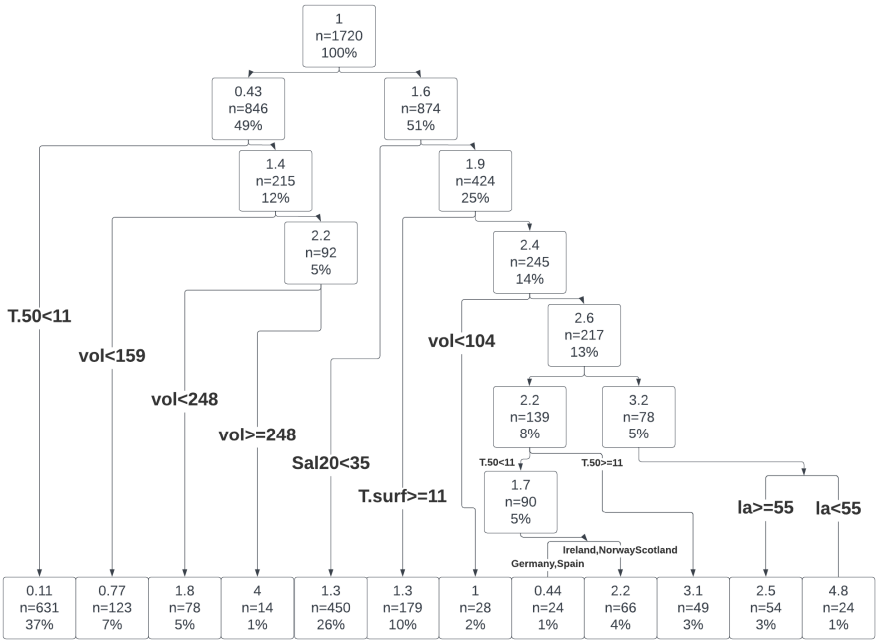

Fig. 3. Tree plot.

Table 2. Grid Search on "mtry" in random forest model based on RMSE on both training set and test set.

mtry	10	20	50	100	150
Test set	1.331	1.125	1.088	1.097	1.102
Train set	1.305	0.861	0.481	0.462	0.461

3.2.4 Projection Pursuit.

To capture non-linearity and improve prediction accuracy, we used a projection pursuit model to explain the more complex relationship between predictor variables and the response variable by a technical report [8]. Projection pursuit models are particularly useful in identifying significant directions in the predictor space that contribute to the variability in the response variable.

Model setup: We used the training set as the data frame for fitting the model. We set 'nterms,' the number of basis functions to include in the model, to 6; 'optlevel,' the optimization level for fitting the model, to 3; and 'max.terms,' the maximum number of terms to be considered during the model fitting process, to 10.

3.3 Cross-Validation

We divided our original dataset into two subsets. We randomly assigned 80% of the data from the original dataset to the training set and the remaining 20% to the test set.

We then applied cross-validation on the training and test sets using the default fold number, $k = 10$. This approach allows us to compute the mean square error (MSE) and R^2 for the training and test sets separately. Consequently, we can compare the performance of estimation and prediction for each model based on these metrics. Additionally, this method helps us detect any overfitting within each model.

4 Results

After fitting the four statistical regression models, we apply cross-validation to evaluate their prediction performance. Additionally, we will discuss feature importance and the differences between the two fish species, as outlined in our research questions.

4.1 Estimation

In Table 3, LASSO regression exhibit the largest MSE and smallest R^2 among the four models, indicating that the dataset likely contains non-linearity. The projection pursuit model performs relatively better, despite its MSE and R^2 values on the training set being 1.275 and 0.446, respectively, its prediction performance is sub-optimal.

Tree-based models, particularly the decision tree and random forest, perform the best among all five models. Both models achieve low MSE on the test set, with the random forest model showing the highest R^2 on the test set at 0.445.

Table 3. MSE on training set and test set of each models.

Model	Train MSE	Test MSE	Train R^2	Test R^2
LASSO regression	1.826	1.719	0.206	0.219
Tree	1.419	1.555	0.383	0.298
Random Forest	0.235	1.221	0.928	0.445
Projection Pursuit	1.275	1.605	0.446	0.293

4.2 Feature Importance

In Table 4, we calculate the feature importances of important features as shown in a technical report [6]. We can see that the feature 'Type' has a %IncMSE of 58.925, indicating that the mean square error would significantly increase if 'Type' were randomly excluded.

The 'Type,' 'T.surf,' and 'T.50' columns in the IncNodePurity measure exhibit relatively high values among all factors, indicating a significant total decrease in node impurity. Consequently, we can conclude that within the top 10 important features, fish type and temperature are the most significant factors affecting fish egg count.

Table 4. Top 10 important features in random forest model.

Feature	%IncMSE	IncNodePurity
Type	58.925002	585.19640
T.surf	21.564896	593.15728
T.50	19.274761	648.24259
Lo	18.607336	424.53753
La	16.433225	371.42137
Sal20	16.368830	478.36831
Vol	15.297500	453.86110
countryIRELAND	8.329785	56.16555
countrySCOTLAND	7.285715	35.03507
GearGULF7	6.163316	30.79391

4.3 Difference in Two Fishes

As illustrated in Table 4, the variable 'Type' is significant to the response variable 'log(count)' in our random forest model.

5 Conclusion

Although the R^2 on the training set is significantly higher, indicating overfitting, we select the random forest model as our final model for dataset estimation due to its superior performance.

The variable 'Type' is the most important feature related to the response variable, 'log(count).' Additionally, sampled water temperatures, the position of the sample station, salinity, and the volume of sampled water are relatively important features that significantly affect the response variable.

We can conclude that there is a difference in the production of horse mackerel eggs and mackerel eggs. Furthermore, since the estimate of 'Type' in our LASSO regression is -1.130, it indicates that the average egg production count of horse mackerel is greater than that of mackerel.

6 Limitations

There are still some limitations within this survey and analysis:

In our regression analysis, we used the dataset from the 'Mackerel and Horse Mackerel Survey' conducted in 2010. Since we are using only a small subset of the original dataset, our conclusions may not be entirely accurate.

We used only five models to fit the entire dataset, leaving many potential models unexplored. Overfitting has caused serious problem in our chosen random forest model as we can see from Table 2. For further analysis, we could try fitting K-Nearest Neighbors (KNN), spline regression, and other machine learning models to see if there are any improvements in the results.

The original dataset contains too many missing values. As a result, we omitted all observations with missing data for any factor, leading to a significant loss of data before fitting our models. If we had more observations with fewer missing values, we could draw stronger conclusions in our regression analysis.

References

1. Wood, S. (2019). Gamair: Data to Accompany Wood (2006) Generalized Additive Models: An Introduction with R, (Version 1.0-2) [Data set: med_meh]. R package.
2. Tay, J.K., Narasimhan, B., and Hastie, T. (2023). "Elastic Net Regularization Paths for All Generalized Linear Models." Journal of Statistical Software, **106**(1), 1-31. doi:10.18637/jss.v106.i01. https://doi.org/10.18637/jss.v106.i01.
3. Liaw, A. and Wiener, M. (2002). "Classification and Regression by randomForest." R News, **2**(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.
4. International Council for the Exploration of the Sea (2024). "Eggs and larvae." https://www.ices.dk/data/data-portals/Pages/Eggs-and-larvae.aspx.
5. R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.
6. Kuhn, M. (2008). "Building Predictive Models in R Using the caret Package." Journal of Statistical Software, **28**(5), 1-26. doi:10.18637/jss.v028.i05. Available at https://doi.org/10.18637/jss.v028.i05.
7. Therneau, T. M. and Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. Available at https://CRAN.R-project.org/package=rpart.
8. Friedman, J.H. and Stuetzle, W. (1981). "Projection Pursuit Regression." Journal of the American Statistical Association, **76**(376), 817-823. doi:10.2307/2287575. Available at https://doi.org/10.2307/2287575.