# Forecasting Mackerel Fishery

**A Spatially Correlated Timeseries Problem**

**Technical University of Denmark**

---

42186 Model Based Machine Learning

**May 30, 2020**

Frederik Bjare, s082633

Kia Kafaei Yahyavi, s153811

Stanley Frederiksen , s140425

## Table of contents

# 1 Introduction

Mackerels are some of the most delicious fish that swims actively in the upper 25-30 fathoms of the water in the summer. During winter, however, they descend to as deep as 100 fathoms. The mackerels spawn during the spring and early summer along the coastlines. Mackerel fishery is one of the most important fisheries in EU with yearly quotas of more than 900.000 tonnes.

We want to analyze the data in the ICES Fishmap database which contains a 30 year long time series of 73.000 samples of no. of mackerels caught at various locations during scientific surveys. Geographically, it spanned most of the oceans and seas of Europe.

## 1.1 Aim of our project

The goal of this project is to make predictions on how many Mackerels will be caught in the future - within various time intervals and different subareas of EU waters - based on the Fishmap-dataset. This could be useful for the fishing industry in able to tell where the likelihood of catching most mackerels would be.

# 2 Methods

Initially, we split the data geographically (by 5 degrees of latitude) as shown in figure 1. The red dots are the locations of the mackerels and the green horizontal lines are the latitude (vertical axis) borders that separates the data set into 4 different areas. However, we have only used the samples between 40 to 60 degrees of latitude, as the data was sparsed outside these boundaries.

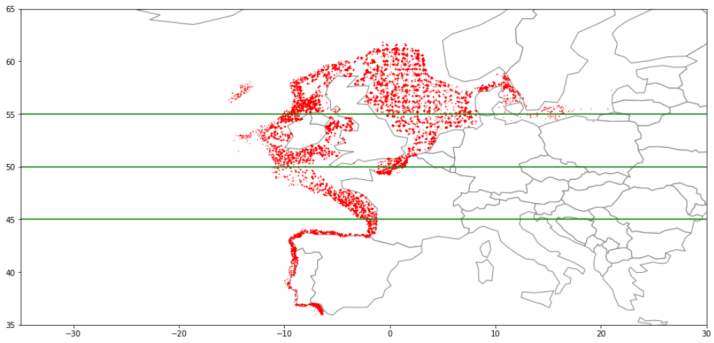


Figure 1: Map of Europe. The red color indicates the location of the mackerels. Green lines are the borderlines that separates each area.

## 2.1 Exploring the data set

Descriptive analysis was performed, producing various plots and histograms that helped in gaining insight in the patterns of the data. Figure 2 shows the time series (mean catch value per year) for each of the 4 areas for the last 30 and 5 years, respectively.

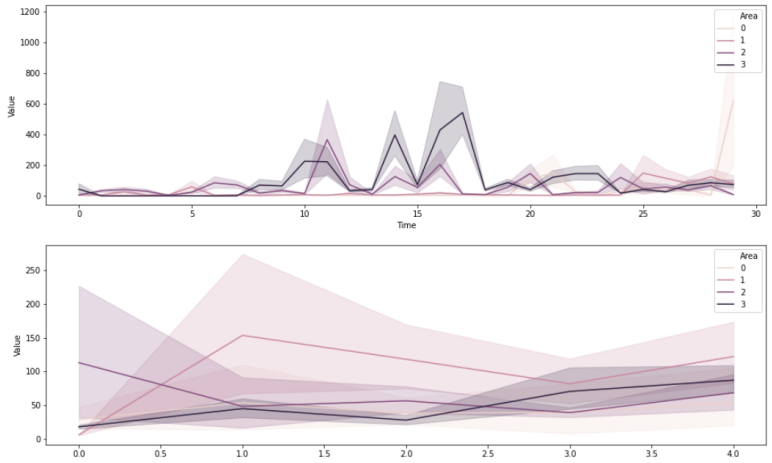


Figure 2: Time series of mackerel catch for the last 30 (upper) and 5 years (lower) for the 4 different areas.

While only weak evidence of autocorrelation was indicated in the dataset (using correllograms and fourrier analysis), we proceeded with a timeseries analysis as this model had already been settled upon.

## 2.2 PGM and generative story

Figure 3 shows our graphical model followed by a generative story, which forms the foundation of our upcoming modeling. The hidden state is defined as:

$$\underline{z_t} = diag(\underline{\beta_1})\ \underline{z_1} + \underline{z_t}\ \underline{\underline{SC}}$$

Matrices are denoted by double underline and vectors by single underline. In the above, $\underline{\underline{SC}}$ is the correlation between the areas and the i'th element of $\beta$ is the linear dependence of the state on previous value in i'th area. This gives us a measure for updating timesteps as:

$$\underline{z_t} = (diag(\underline{\beta}_1)\underline{z_{t-1}})(\underline{\underline{1}} - \underline{\underline{SC}})^{-1}$$

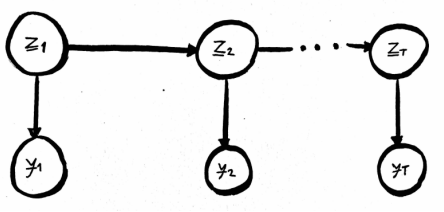

Figure 3: PGM where:
$\underline{\mu}_{z_t} = (diag(\underline{\beta}_1)\underline{z_{t-1}})(\underline{\underline{1}} - \underline{\underline{SC}})^{-1}$
$\underline{z_t} \sim Normal(\underline{z_t}|\mu_{z_t}, \tau)$
$\underline{y_t} \sim Normal(\underline{y_t}|\underline{z_t}, \sigma)$

**Generative story**

Draw dependence on past states $\underline{\beta_1} \sim Normal(\underline{\beta_1}|0, 1)$

Draw correlation between present states $\underline{\underline{SC}} \sim Normal(\underline{\underline{SC}}|\underline{\underline{0}}, \underline{\underline{1}})$

Draw state uncertainty $\tau \sim Normal(\tau|0, 1)$

Draw observation noise $\sigma \sim Normal(\sigma|0, 1)$

For each time $t \in \{1, 2, ..., T\}$

Calculate means of distributions to draw from:

$\underline{\mu}_{z_t} = (diag(\underline{\beta}_1)\underline{z_{t-1}})(\underline{\underline{1}} - \underline{\underline{SC}})^{-1}$

For each area $a \in \{1, 2, ..., A\}$

Draw hidden state $\underline{z_t} \sim Normal(\underline{z_t}|\underline{\mu_{z_t}}, \tau)$

Draw observation $\underline{y_t} \sim Normal(\underline{y_t}|\underline{z_t}, \sigma)$

### 2.3 Models

Different models were applied using both Pyro and STAN. We approached our modeling by the following steps:

1. Global mean (baseline)
2. First-order autoregressive
3. Resampling - Seasonal based (3 months interval)
4. Multivariate LDS

where each of the model was evaluated in terms of errors to see the performance. All models were fitted for a 30-year period (yearly interval) except the resampling model (19 years with a quarterly interval).

# 3 Results

In general the models didn't do a fantastic job at predicting compared to our own expectations. Table 1 shows the results of the different models.

| Performance metric | Mean prediction | Autoregression | Resampling | Spatial correlation and Imputation |
|---|---|---|---|---|
| MAE | 27.82 | 20.71 | 60.69 | 20.19 |
| RAE | 1.41 | 1.06 | 0.87 | 1.03 |
| RMSE | 34.28 | 28.93 | 147.46 | 28.82 |

Table 1: Model evaluation by the Mean absolute error (MAE), Realtive absolute error (RAE) and Root Mean Square Error (RMSE)

The spatial correlation and imputation of the future values seems to have the best performance of all the models compared to our baseline in terms of MAE and RMSE. Although, only with a minor improvement in contrast to the autoregressive model. The resampling method had the lowest RAE which actually makes it the best predicting model. This might be due to the fitting intervals as predicting for a shorter period (3 months) compared to a whole year can be more accurate because of lesser variation between the intervals.

# 4 Conclusion and future work

Using the global mean for predicting future values may be inaccurate. One may also consider the trade-off between computational complexity and the improvement of the models (i.e. AR may be sufficient enough for prediction). Modeling in different (shorter) time intervals seems to have a better forecasting accuracy as seen with the resampling method.

## 4.1 Using External variables

Given more time a few things would have been tried out. One would be using external variables. To do that properly some time would've had to be spent doing feature engineering and acquiring domain knowledge. Basically researching what kind of factors play a role in hauling fish from the ocean.

### 4.2 Different distributions for output

After implementing the final iteration of the models further descriptive analysis was performed. Figure 4 shows the distribution of fish, caught at different time steps, in different areas. Essentially, it shows that when a ship goes out to catch mackerels, a very large amount of the time when they cast their net, only one fish is caught (0 catches are not registered). On the contrary, very few times, a very large amount of fish are caught ($10^3$ or $10^4$ fish). It seems that the distribution of fish caught follows something similar to an exponential distribution.

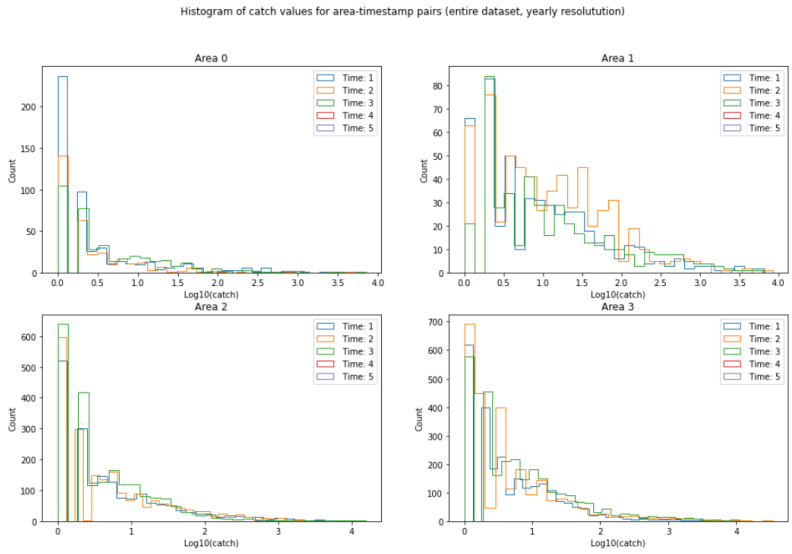


Figure 4: Catch values

Unfortunately, the fact that the data might be exponentially distributed was realized a bit too late, and given more time it would be interesting to try out different distributions for the output.