**Problem Chosen**
A

**2020**
**MCM**
**Summary Sheet**

**Team Control Number**
2014190

# Moving North: Effect of Increased Water Temperature on the Habitat Preference of Fish

## Summary

Fishery is one of the most important industry of Scotland. But according to many researches in the recent years, fish like herring and mackerel is moving their habitats northwards. One of the biggest reason is the climate change, especially the sea temperature.

Therefore, in this paper, our team is trying to predict where the two species of fish, herring and mackerel could be over 50 years, which can make fishermen realize what they should do in the future.

For starters, we predict the environment factors that can influence fish's habitat such as sea surface temperature and sea surface salinity. We learned that based on Ensemble Empirical Mode Decomposition(EEMD) method and Autoregresive Integrated Moving Average(ARIMA) model, we can predict them more precisely. We get a great amount of history data from official institutes such as National Oceanic and Atmospheric Administration(NOAA),so we can use the model efficiently. And the result shows that the sea surface temperature around Scotland will become a little higher in the future.

Second, the most import part, is to predict the fish location by using the history and future environment data. We get the fishery data from database of International Council for the Exploration of the Sea(ICES).After testing the statistics indexes for different prediction models, we choose on Generalized Additive Model(GAM) to obtain our result, as the model has flexible predictor functions which can uncover hidden patterns in the data, and its auto-regularization of predictor functions can help to avoid over-fitting. It can deal with many parameters and turn them into an ideal prediction model. The result shows that the cluster of fish is moving northwards which meets the expectations. Based on the ecological predicted data generated by EEMD and ARIMA model, we view it as the test data and feed it into our GAM model, then the fish density distribution data is predicted.

Finally, as a conclusion, we figure out that herring and mackerel will move to the north place substantially if the sea temperature in the future increases higher and higher. Although our result shows there won't be a big impact on the fishery in the next 50 years, we need to be cautious about the rising trend of temperature. Otherwise, fishermen need to explore more further in the sea and better vessels are required in the future.

**Keywords**: Fish location prediction, Ensemble Empirical Mode Decomposition, ARIMA, Generalized Additive Model.

# Contents

# 1 Introduction

## 1.1 Background

Fishing is a big business, culturally important, and supports a significant number of jobs in Scotland. Each year, about 446 thousand tons of seafood are harvested [1], contributing much million to GDP in fishing industry and across the broader economy. Among them one third are herring and mackerel. Oceans have provided enormous benefits to the Scottish.

(a) Herring

(b) Mackerel

Figure 1: Illustration of the of herring and mackerel, Source: *Wikipedia*

However, climate change is compromising these benefits by modifying the stock location. In particular, global ocean warming has altered the distribution of several marine fish species. The species have shifted northward to follow the displacement of their thermal habitat, which is known as *niche tracking*[2].

The niche tracking may have some potential implications on local ecosystems, breaking the established food chain and affecting fisheries directly. Hence, establish a reliable projections of future changes in fish distribution based on history data is strongly needed.

## 1.2 Restatement of the Problem

We are required not only to develop a model related to the migration of Scottish herring and mackerel to predict their future distribution, considering the ocean temperature, but also to identify strategies for small fishing companies.

Our tasks is to:

- build a model describing the change of sea surface temperature around Scotland, given a specified grid and date with several parameters unsettled.
- build another model to show the relationship between the distribution of the given two species and the sea surface temperature.
- adjust parameter in the temperature model to indicate some possible scenarios, predict the best/worst case and most likely elapsed times until those small fishing companies are not able to harvest without change current locations
- provide strategies for fishing companies based on our prediction

# 2 Assumptions and Notations

## 2.1 General Assumptions

Our model make the following assumptions:

- **Historical data used in our model are valid and accurate.** To solve this problem, data from ERDDAP provided by National Oceanic and Atmospheric Administration(NOAA) is used in our model. We assume that the historical data provided are accurate. Besides we ignore the absense of some data.
- **The marine area of the Scotland will not change in the next 50 years.** The permitted marine area changed several times in the history, we can not predict how the world changed over the next 50 years, so we assume that the future marine area is the same as today in Scotland.
- **The small fishing companies can be found everywhere along the coastline.** The full data about those small fish communities and single persons are extremely hard to find, we make them everywhere to simplify the problem. Also we assume that the tools they used can support at most a 150-miles round trip to catch the fish.
- **The future is predictable.** That is to say, in the model, we assume that the inuence of exogenous factors can be ignored over time.

## 2.2 Notations

| Notation | Description |
|---|---|
| $p$ | The order of the autoregressive model |
| $q$ | The order of the moving-average model |
| $c_n$ | The IMF decomposed from the $n_{th}$ EEMD |
| $r_n$ | The remain data after n times EEMD |
| $N_k$ | Total population in age group $k$ |
| $SSS$ | Data of sea surface salinity ($\subseteq \mathcal{R}^n$) |
| $SST$ | Data of sea surface temperature ($\subseteq \mathcal{R}^n$) |
| $SSH$ | Data of sea surface height from seabed ($\subseteq \mathcal{R}^n$) |
| $LON$ | Data of longitude grid scope in the area($\subseteq \mathcal{R}^n$) |
| $LAT$ | Data of latitude grid scope in the area($\subseteq \mathcal{R}^n$) |
| $DoF$ | Data of density distribution of each fish($\subseteq \mathcal{R}^n$) |
| $\alpha$ | The constant of fitted model |

Table 1: Notations in our model

# 3 Models

## 3.1 Setting Up Grid

Setting up geogrid is the most common technique when we deal with data associated with location. In our model, the grid with $1^\circ \times 1^\circ$ latitude-longitude resolution is used, dividing the map into equal-area squares.(**Figure 2**)

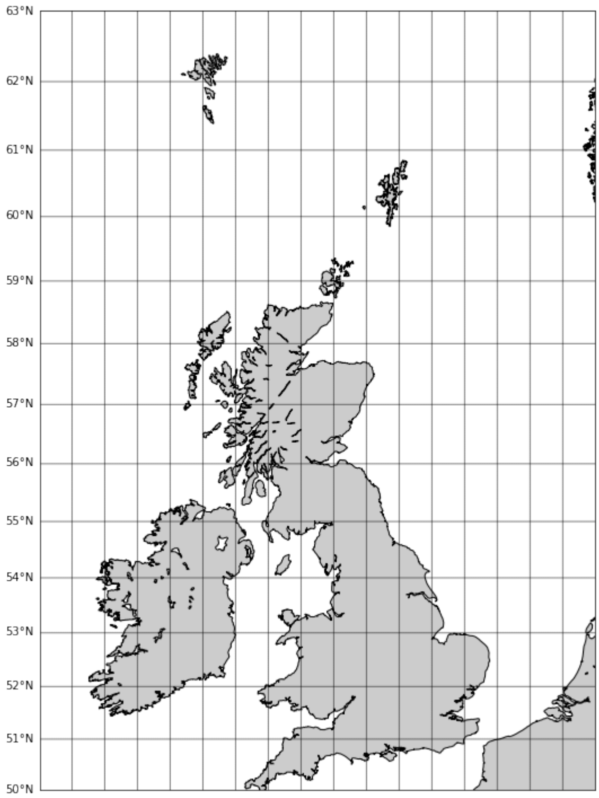


Figure 2: Example Grid of the United Kingdom

## 3.2 Environment Data Prediction Based on Ensemble Empirical Mode Decomposition and Auto-Regressive Integrated Moving Average

The first problem for predicting the fish location is to get the future sea surface temperature($SST$), sea surface salinity($SSS$) and surface height from seabed($SSH$). As the improvement and development of science and technology, we can get plenty of the history data. Therefore, we can use Ensemble Empirical Mode Decomposition (**EEMD**) to decompose our data of different years into different frequency series, and then use Auto-Regressive Integrated Moving Average (**ARIMA**) model to predict them respectively, and compose the results into the final result, then we can get a not bad prediction.[3]

### 3.2.1 Ensemble Empirical Mode Decomposition (EEMD)

EEMD is a way to decompose a signal into the so-called intrinsic mode functions(IMF) along with a trend, and obtain instantaneous frequency data[4]. It is based on Empirical Mode Decomposition(EMD). The basic method of EMD is to take the envelope defined by the extreme points to obtain the result. At first, draw two extreme point envelopes of $y(t)$, and then compute the average envelopes. Assuming that the average is $m_1$, the difference of $y(t)$ and $m_1$ is $h_1$:

$$h_1 = y(t) - m_1 \tag{1}$$

But in most conditions, $h_1$ cannot fit the definition of IMF directly, so we need to deal with $h_1$. There we make $h_1$ as a new record and use the method above to get another average of envelopes, and we name it as $m_{11}$. Then we get the difference of $h_1$ and $m_{11}$:

$$h_{11} = h_1 - m_{11} \tag{2}$$

And Then we repeat the operation $k$ times until $h_{1k}$ can fit the definition of IMF, we denote:

$$c_1 = h_{1k} \tag{3}$$

There $c_1$ is our first decomposed IMF, and it contain the shortest cycle component of the whole signal data. After we decompose $c_1$, we can get:

$$r_1 = y(t) - c_1 \tag{4}$$

Obviously, there are many long cycle components in the remain $r_1$. Therefore, we make $r_1$ as the new data and repeat the operation above and get the final results:

$$r_2 = r_1 - c_2, \cdots, r_n = r_{n-1} - c_n \tag{5}$$

The decomposition will end when $r_n$ becomes a constant, a monotonic function or a function with only an extreme point. We called $r_n$ the remain. Combine the equations above we get:

$$y(t) = \sum_{i=1}^{n} c_i + r_n \tag{6}$$

Using the EEMD method above, we can decompose our SST data(SSS, SSH similarly) of 36 years(1985-2020) into time series of different frequency. Then we can use the ARIMA model to predict future data.

### 3.2.2 The Auto-Regressive Integrated Moving Average (ARIMA) Model

An Auto-Regressive Integrated Moving Average (ARIMA) model is a generalization of an Auto-Regressive Moving Average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.[5]

We use this model to predict future environment data due to the time series produced by EEMD. The mathematical form of general ARMA model is:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \cdots + \varphi_q \varepsilon_{t-q} \tag{7}$$

where $\phi_1, \phi_2, \cdots \phi_p$ are the parameters of the auto-regressive part, $\varphi_1, \varphi_2, \cdots \varphi_q$ are the parameters of the moving average part, $\varepsilon_t$ is the white noise sequence. The formula is denoted as $ARMA(p,q)$.

To use this model, we should justify the stationarity of our original environment data and the IMF series after EEMD. Then we do the order determination. We use BIC which is based on *Akaike information criterion*[6]. For $ARMA(p,q)$, the BIC function is:

$$\mathrm{AIC}(p,q) = n \ln\left(\hat{\sigma}^2\right) + 2(p+q)\ln n \tag{8}$$

Finally, according to the prediction formula of $ARMA(p,q)$:

$$\hat{X}_k(l) = \begin{cases} \hat{X}_k(l) & l \geqslant 1 \\ \hat{X}_{k+l}(l) & l \leqslant 0 \end{cases} \tag{9}$$

The variance is:

$$\mathrm{Var}\left[e_k(l)\right] = \left(1 + G_1^2 + \cdots + G_{l-1}^2\right)\sigma_e^2 \tag{10}$$

where the function G is:

$$\begin{aligned}
&G_0 = 1 \\
&G_l = \sum_{j=1}^{l} \phi_j^\star G_{l-j} - \theta_j^\star \quad (l = 1, 2, \cdots q) \\
&\phi_j^* = \left\{ \begin{array}{ll} 0 & j > p \\ \phi_j & j \leqslant p \end{array} \right. \\
&\theta_j^* = \left\{ \begin{array}{ll} 0 & j > p \\ \theta_j & j = 1, \cdots, q \end{array} \right.
\end{aligned} \tag{11}$$

The prediction vector is $\hat{X}_k = \left(\hat{X}_k(1), \hat{X}_k(2), \cdots \hat{X}_k(l)\right)^{\mathrm{T}}, l = 1, 2, \cdots q$. And the recursive formula can be turned to:

$$\hat{X}_{k+1} = \phi_1 \hat{X}_{k+1}(l-1) + \phi_2 \hat{X}_{k+1}(l-2) + \cdots + \phi_p \hat{X}_{k+1}(l-p) \quad (l > p) \tag{12}$$

#### 3.2.3 Prediction

Using the EEMD method, we can decompose our environment data within the scope of location(40°N~65°N, $-20$°W~5°E) of 35 years(1985~2020) into different IMF. Next, predict the future environment data under the ARIMA model of every IMF. Finally, re-compose all the components into our prediction result.

### 3.3 Fish Distribution Prediction Based on Generalized Additive Model

To solve this issue, let's introduce the mathematics model Generalized Additive Model, which is known as GAM. GAM was originally developed by *Trevor Hastie and Robert Tibshirani*[7] to blend properties of generalized linear models with additive models. Despite its lack of popularity in the data science community, GAM is yet a powerful simple technique, based on its strong mathematics theory. Here, we will briefly give out some mathematics introduction and a simple and explicit example on GAM's application on the relationship between a kind of wine's features and its quality score, thus you can understand this model better. Finally we will give out our understanding about this problem and GAM's application on solving this problem.

#### 3.3.1 The Mathematics Theory of Generalized Additive Model

In statistics, a generalized additive model(GAM) is just like a Generalized Linear Model(GLM) but more than a GLM model. Detailed introduction and mathematics theory can be found on wikipedia[8]. Mathematically speaking, GAM is an additive modeling technique where the impact of the predictive feature variables is captured through smooth functions which is depending on the underlying patterns behind the data. Because its generalization extended by smooth functions, it has some merits in model fitting:

- Easy to interpret
- Flexible predictor functions can uncover hidden patterns in the data
- Regularization of predictor functions help to avoid over-fitting

In general, GAM has better interpretability advantages of GLMs where the contribution of each independent feature variable to the prediction target variable is clearly encoded. However, it substantially has more flexibility because the relationships between feature variables (independent) and target variables(dependent) are not assumed just to be linear or simply polynomial. Actually, we dont have to know a priori which type of predictive functions we will eventually need. From an estimation standpoint, the use of regularized, non-parametric functions avoids the pitfalls of dealing with higher order polynomial terms in linear models. From an accuracy standpoint, GAMs are competitive with other popular learning techniques and models. Therefore, many biologists and ecologists take it into their arsenal to assist them with the research process of finding some underlying pattern among a large quantity of different kinds of data in various features. During the process of solving this problem, we test the effect on the same ecosystem dataset with some popular learning models, like linear regression models and polynomial regression models, and finally choose GAM because of its better performance.

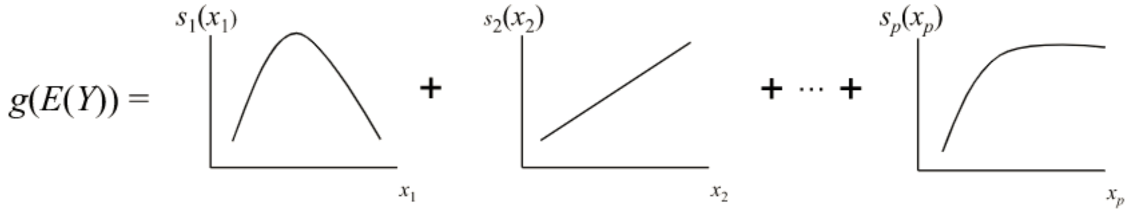


Figure 3: A Simple Visualization of GAM Mathematical Theory

Also we can write the model structure as:

$$g(E(Y)) = \alpha + s_1\left(x_1\right) + \cdots + s_p\left(x_p\right) \tag{13}$$

where $Y$ is the dependent variable (i.e., what we are trying to predict), E(Y) denotes the expected value, and $g(Y)$ denotes the link function, or the transform function, which links the expected value to the predictor feature variables $x_1, \ldots, x_p$. The terms $s_1\left(x_1\right), \ldots, s_p\left(x_p\right)$ denote those smooth, non-parametric functions. Note that, in the context of regression models, the terminology non-parametric means that the shape of predictor functions are fully determined by the data as opposed to parametric functions that are defined by a typically small set of parameters. This can allow for more flexible estimation of the underlying predictive patterns without knowing upfront what these patterns look like.

In the aspect of interpretability, it's easy to attain some reliable message from GAM model. When a regression model is additive, it means that the interpretation of the marginal impact of a single variable (the partial derivative) does not depend on the values of the other variables in the model. Hence, by simply looking at the output of the model, we can make simple assertions about the effect of the predictive variables that make sense to a non-technical person. For instance, for the visualization graphic illustration above, we can say that the transformed expected value of $Y$ increases linearly as $x_2$ increases, holding other independent variables constant. Or, the transformed expected value of $Y$ increases with $x_p$ until $x_p$ hits a certain point, etc.

Additionally, an important feature of GAM is the ability to control the smoothness of the predictor functions. With GAMs, you can avoid wiggly, nonsensical predictor functions by simply adjusting the level of smoothness. In other words, we can impose the prior belief that predictive relationships are inherently smooth in nature, even though the dataset at hand may suggest a noisier relationship. This plays an important role in model interpretation as well as in the believability of the results.

In the aspect of flexibility and automation, GAM can capture common nonlinear patterns that a classic model would miss. For example, the genuine pattern between one feature variable $x_i$ and the response variable $Y$ can be of curve shape, but many linear models would miss this important curve feature and regard it as a linear relation.

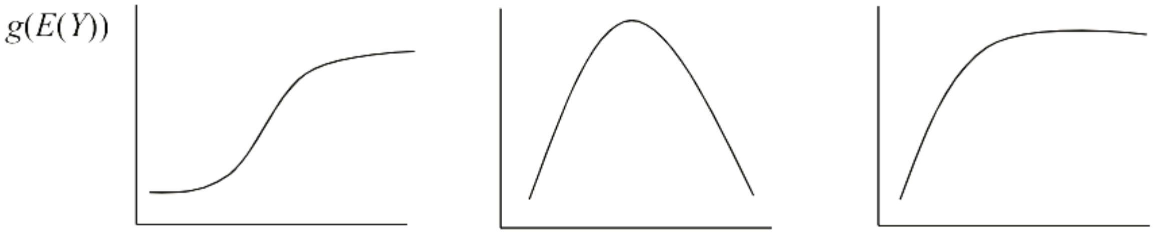


Figure 4: "Mountain-shaped" Patterns between each Feature Variable and target variable $Y$

When fitting some parametric regression models, these non-linear effects can be typically captured via binning or polynomials. This leads to clumsy model formulations with many correlated terms and counterintuitive results. Moreover, selecting the best model involves constructing a multitude of transformations, followed by a search algorithm to select the best option for each predictor a potentially greedy step that can easily go awry. Hence, We dont have this problem with GAM. Predictor functions are automatically derived during the model estimation. We dont have to know upfront what type of functions we will need. This will not only save us time, but will also help us find patterns we may have missed with a parametric model.

Finally, in the aspect of GAM's ability to avoiding over-fitting, a smoothing parameter $\lambda_i$ is used to control smoothness of the predictor functions.And this smoothing parameter $\lambda_i$ can be trained to attain its best value by grid-search, which is a machine learning training method. By controlling the wiggliness of the predictor functions, we can directly tackle the bias/variance trade-off. Moreover, the type of penalties applied in GAM have connections to Bayesian regression and $l2$ regularization.

To fit this model and avoid over-fitting, a penalized likelihood score function is applied in this model. For both local scoring and the GLM approach, the model's ultimate goal is to maximize the penalized likelihood function, although they take very different routes. The penalized likelihood function is given by:

$$2l\left(\alpha, s_1\left(x_1\right), \ldots, s_p\left(x_p\right)\right) - penalty \tag{14}$$

where $l\left(\alpha, s_1, \ldots, s_p\right)$ is the standard log likelihood function. For a binary GAM with a logistic link function(or transform function), the penalized likelihood is defined as below:

$$\begin{aligned} l\left(\alpha, s_1\left(x_1\right), \ldots, s_p\left(x_p\right)\right) = \textstyle\sum_{i=1}^{n}\left(y_i \log \hat{p}_i + \left(1-y_i\right) \log\left(1-\hat{p}_i\right)\right) \\ \hat{p}_i = \left(1+\exp\left(-\hat{\alpha} - \textstyle\sum_{j=1}^{p} s_j\left(x_{ij}\right)\right)\right)^{-1} \end{aligned} \tag{15}$$

where $\hat{p}_i$ is given as:

$$\hat{p}_i = P\left(Y=1 | x_1, \ldots, x_p\right) = \left(1+\exp\left(-\hat{\alpha} - \sum_{j=1}^{p} s_j\left(x_{ij}\right)\right)\right)^{-1} \tag{16}$$

The penalty can, for example, be based on the second derivatives:

$$\text{penalty } = \sum_{j=1}^{p} \lambda_j \int \left(s_j^{\prime\prime}\left(x_j\right)\right)^2 dx \tag{17}$$

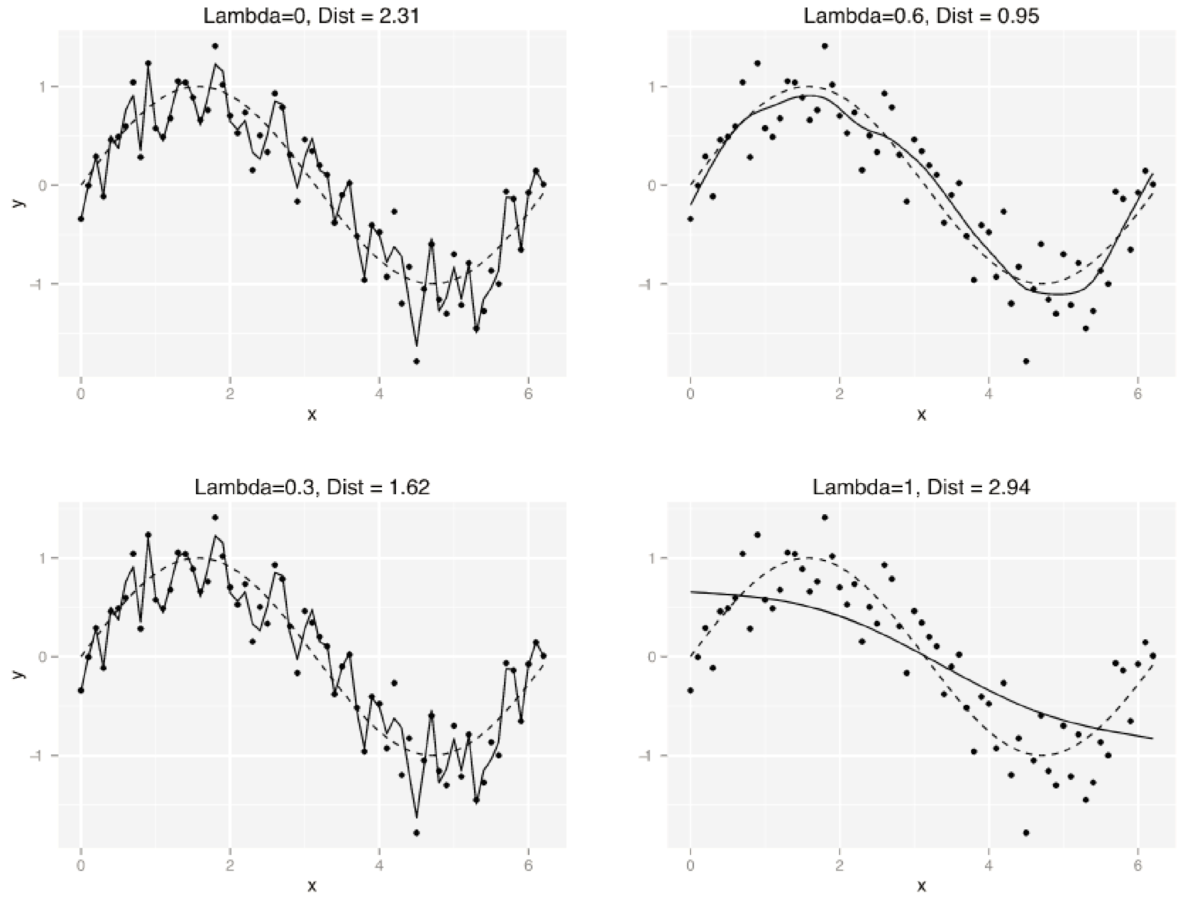
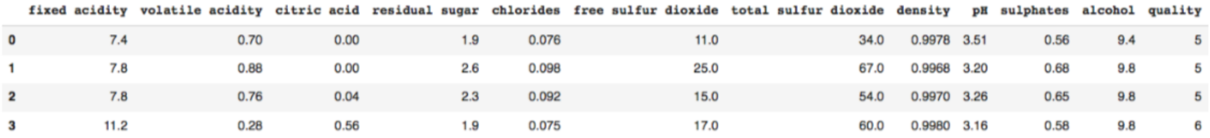


Figure 5: The Effect of $\lambda$ in the Case of Sin(x) Plus Gaussian Noise

The parameters, $\lambda_1, \ldots, \lambda p$, are the aforementioned smoothing parameters which control how much penalty (smoothness) we want to impose on the model. The higher the value of $\lambda_j$, the smoother the curve. These parameters can be pre-selected or trained from the data. Intuitively, this type of penalty function makes sense: the second derivative measures the slopes of the slopes. This means that wiggly curve will have large second derivatives, while a straight line will have second derivatives of 0. Thus we can quantify the total wiggliness by adding up the squared second derivatives.

### 3.3.2 A Simple Example Based on Generalized Additive Model

To officially apply GAM into our task, let's use a simple example to illustrate it. This dataset is about some features and quality scores of the Portuguese Vinho Verde wine, available from the UCI machine learning repository. Input features are 11 physicochemical variables , which describe the red wine qualities from various aspects. The target feature is the quality level score, ranging from 0 to 10, which indicates how good this red wine is.



| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 1 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 2 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| 3 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |

Figure 6: A Brief Overview about UCI Wine Dataset

By training this GAM model with this train-set, some underlying data pattern can be attained as below. Then we build a linear GAM that could predict the red wine quality score based on the physicochemical feature variables. More importantly, how each of these physicochemical feature variables affects the quality score is revealed in the partial dependence plots. For example, as shown above, we know the probable best alcohol percentage is about 13%. And in the pH aspect, we can infer that the more pH index is, the lower the quality score is.

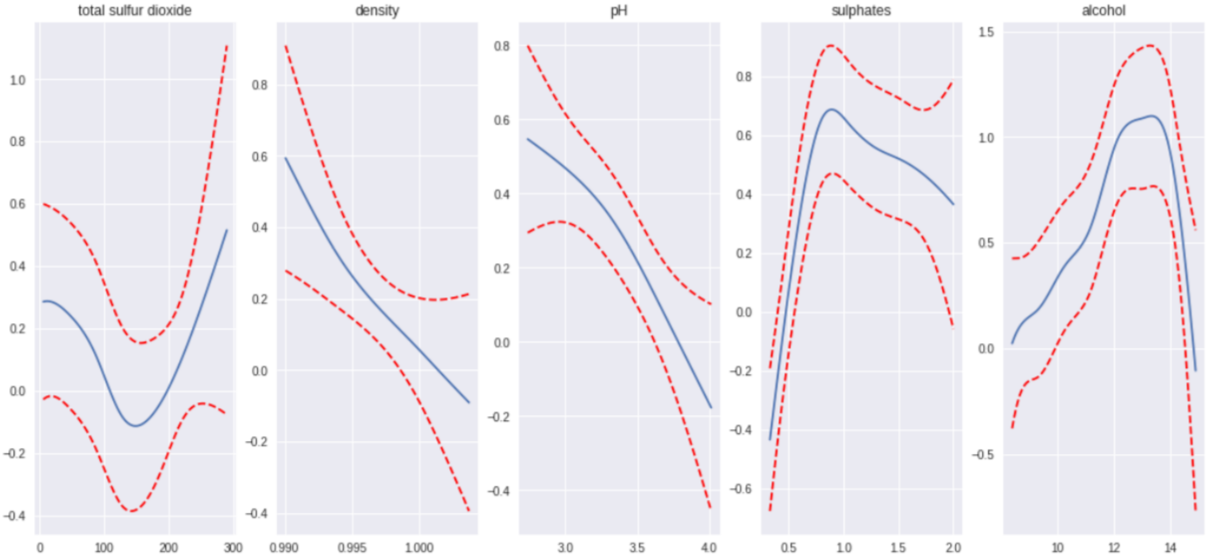


Figure 7: Partial Dependence Plots Showing Some Factors That Affect Wine Quality

By having the underlying pattern, if another batch of wine's feature data is fed into the GAM model, correspondingly, the model can give a batch of believable predictive quality score. That's exactly the way we solve this herring and mackerel distribution prediction problem.

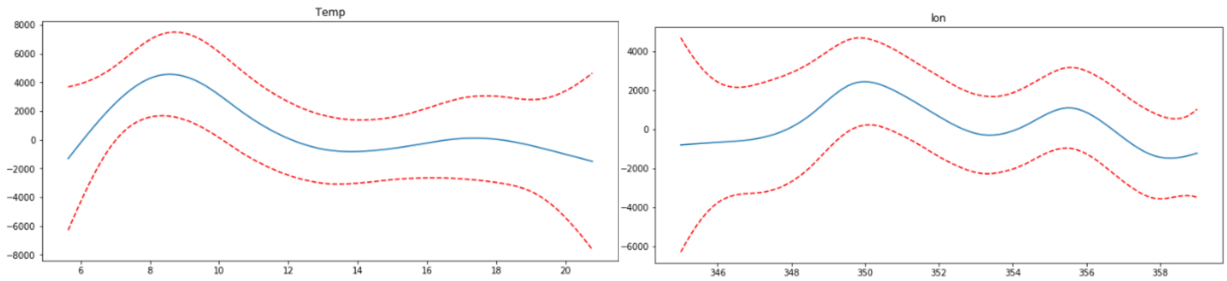
### 3.3.3 Solution to Fish Distribution Prediction Based on Generalized Additive Model

For the first question, to solve our genuine problem, to predict the fish distribution over the 50 years, we apply the GAM model. Actually, it takes quite long time for us to get the data that we need. After researching some biology research articles and papers, we download those ecosystem data from official websites[9][10]. Then, we use Python and its package Pandas to pre-process the data, which includes the monthly sea surface temperature($SST$), the sea surface height ($SSH$), and the sea surface salinity($SSS$). In consideration of those fishes' thriving habitats, we also view the longitude($LON$) and latitude($LAT$) as two feature variables. Our target is to predict the density distribution of fish ,which is shorted as $DoF$(Density of Fish). And actually, the two geographic factors do have a delicate relation with the fish's distribution probability. Now we have the history data of sea surface temperature, salinity, height of different longitudes and latitudes ($-20°$W~$5°$E, $40°$N~$65°$N). We view this five kinds of data as feature variables $x_1, \ldots, x_5$. And we regard the density of herring and mackerel as the target variable $Y$. Before we officially take GAM as the final model, we test the traditional Linear Regression Model and Polynomial Regression Model, but they perform not that good. That is, GAM's fit-effect is much better than these two models, which make sense.For instance, it's not that possible to build a simple linear or polynomial relation between temperature and fish density distribution. So we train this model with the history data in GAM, trying to figure out the underlying patterns behind those data. The integral algorithm can be found on Github[11], which is contributed and stared by many computer scientists. The relation among those data can be expressed in the equation below.

$$g(E(DoF)) = \alpha + s_1\left(SSS\right) + s_2\left(SST\right) + s_3\left(SSH\right) + s_4\left(LON\right) + s_5\left(LAT\right) \tag{18}$$

We initialize the $\lambda$ in Gaussian distribution, and train this smoothing parameter in grid-search method. When the training process is finished, we can get the underlying pattern behind these data, which means we can find the actual intercept constant $\alpha$, $s_1(SSS), \ldots, s_5(LAT)$ smoothing functions. Then we put the future data that we have generated by EEMD and ARIMA models into the GAM model, then the density distribution of these two types of fish can be predicted. And we have some findings.

As the picture shows, we can learn that mackerels prefer to live in the temperature scope in $7°C$~$10°C$ environment, while herrings don't have a conspicuous favor to the temperature. Also, the geographic position has something to do with the fish density distribution, as the picture shows, we can acknowledge that the mackerel fish favors to live within the selected scope in longitude.

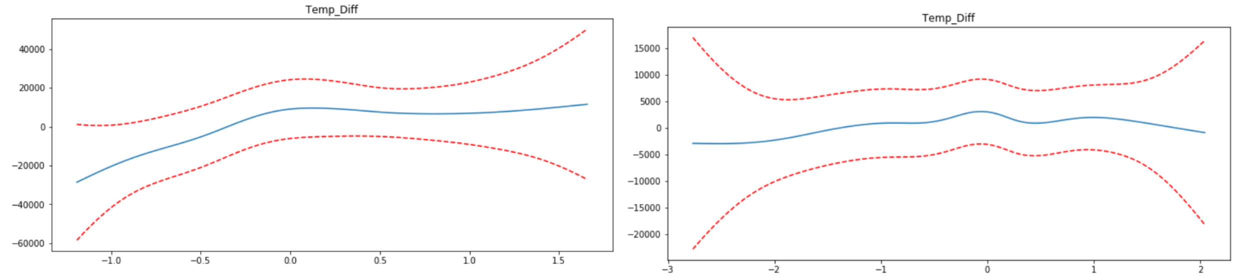


(a) Temperature (b) Longitude

Figure 8: The Underlying Pattern of Some Factors for Mackerel Model

Table 2 is the table for the statistics indexes for GAM in the first problem.

For the second question, we assume that herring and mackerel can live in various temperatures ranged in a rational scope, which means the actual temperature itself will not be the factor that causes the change of fish density distribution. But it's the difference(or the rapidity of the temperature change) between two years that causes the change of fish density distribution. So we use database operations among different data tables to calculate the speed of temperature change in the same position each two years.This time, we use SSS, SSH, LON, LAT, and the temperature difference between two years on sea surface as our model's input data, then we train this GAM model, finding the smoothing functions behind those input feature data along with their corresponding $\lambda_1, \ldots, \lambda_5$. Finally we put the future environment data as feature input data into this trained GAM model, then we get the density distribution of each fish , which we are trying to predict. And here are part of our findings about the second problem. As the pictures shown, we can acknowledge that the mackerel can bear a larger temperature change scope($-3°C$~$+2°C$) than the herring($-1°C$~$+1.5°C$).



(a) Herring (b) Mackerel

Figure 9: The Underlying Pattern of Temperature Change for the Two Models

**Table 3** for the statistics indexes for GAM in the second problem.

| Model | Dist Effective DoF | Link Log Likelihood | AIC | Pseudo R-Squared |
|---|---|---|---|---|
| Mackerel | 31.4252 | -44827.0224 | 89718.8951 | 0.374 |
| Herring | 28.2699 | -24983.795 | 49936.13 | 0.34 |

Table 2: Statistics Indexes for GAMs in the First Problem

| Model | Dist Effective DoF | Link Log Likelihood | AIC | Pseudo R-Squared |
|---|---|---|---|---|
| Mackerel | 22.918 | -11146.8721 | 22341.5801 | 0.517 |
| Herring | 15.4014 | -15430.2863 | 31190.3754 | 0.4221 |

Table 3: Statistics Indexes for GAMs in the Second Problem

# 4 Results

To solve the given problems, the following estimation from our model is given:

- Prediction of sea surface temperature in the selected scope.
- Prediction of the distributions of the herring and mackerel.
- The impact of the migration of the fish on the small fish companies and their strategies.

As mentioned above, historical sea surface temperature used in both predictions is extracted from NCEP Global Ocean Data Assimilation System (GODAS), which provides a monthly data on the $0.333°\sim1°$ latitude–longitude grid. Needed transformation is performed to match the $1°\sim1°$ latitude-longitude resolution in our model. Apart from the temperature data, Sea Surface Salinity (SSS), Sea Surface Height (SSH), and Mixed Layer Depth (MLD) is used when we predict the distributions, such predicted data are extracted from the *CMIP5* database using HadGEM2[12] model by Met Office at RCP 4.5 scenario.

## 4.1 Prediction of Sea Surface Temperature

We predicted Sea Surface Temperature over the next 50 years using the model above. The average SST change in the selected range from 2020 to 2070 is shown in **Figure 10** . From the chart we can easily find that, although the data show a kind of periodicity, the overall trend is always getting warmer and warmer. We believe that the periodicity comes from the change in solar intensity, which is not formally considered in our model.

The temperature projection of the select area in 2025, 2040, 2055 and 2070 is shown in **Figure 11** and **Figure 12** . From the figures, we can easily feel the red zone is expanding north, that is to say, the ocean water is going much warmer than today in the future if the climate change continues.

## 4.2 Prediction of the Distribution of the Species on Problem 1

Using data from multiply source and our predictions of SST and other factors, we now make predictions on the future distribution of the two species.

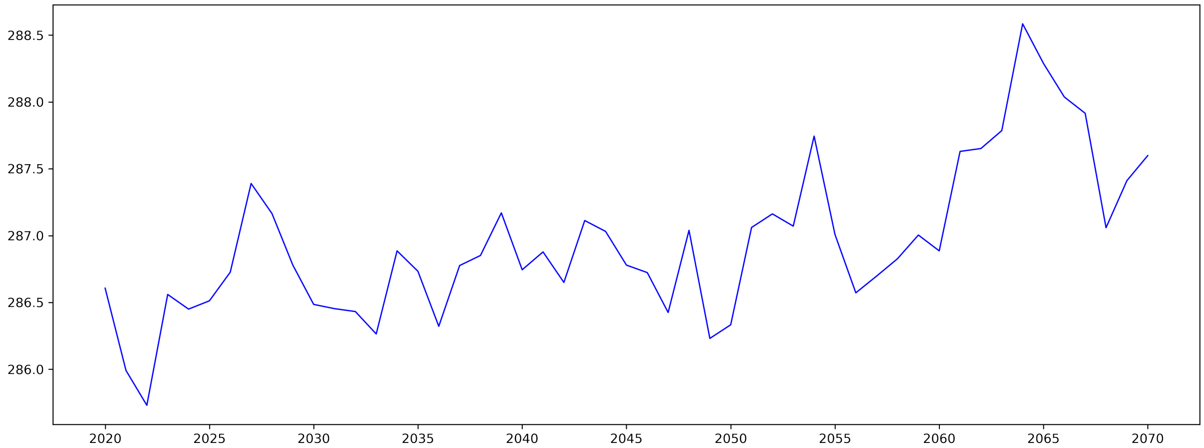
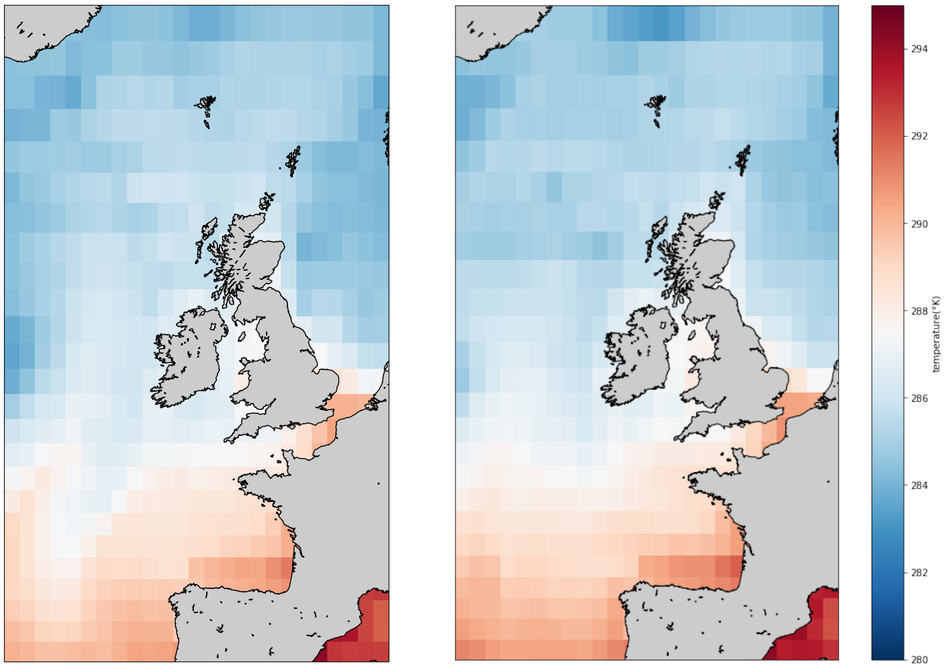


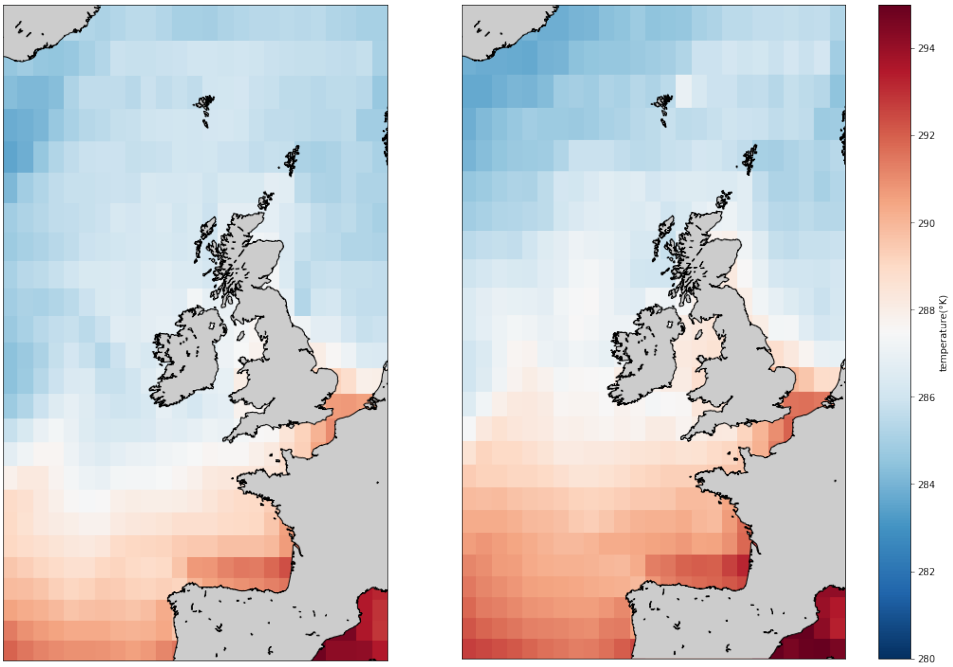
Figure 10: The Estimated Average Sea Surface Temperature (in Kelvin) in the Selected Area



(a) SST in 2025

(b) SST in 2040

Figure 11: The Estimated Sea Surface Temperature (in Kelvin) of the Selected Area



(a) SST in 2055

(b) SST in 2070

Figure 12: The Estimated Sea Surface Temperature (in Kelvin) of the Selected Area (Cont.)

### 4.2.1 Herring

According to our predictions on the density distribution of the two species, we find the main biological community of each kind is geographically moving north, and we name the position of the main community by center of gravity. A Huge latitudinal shift in center of gravity (CoG) of herring and mackerel is measured in our model, as shown in **Figure 13** . We can conclude that SST has a great effect on herring.

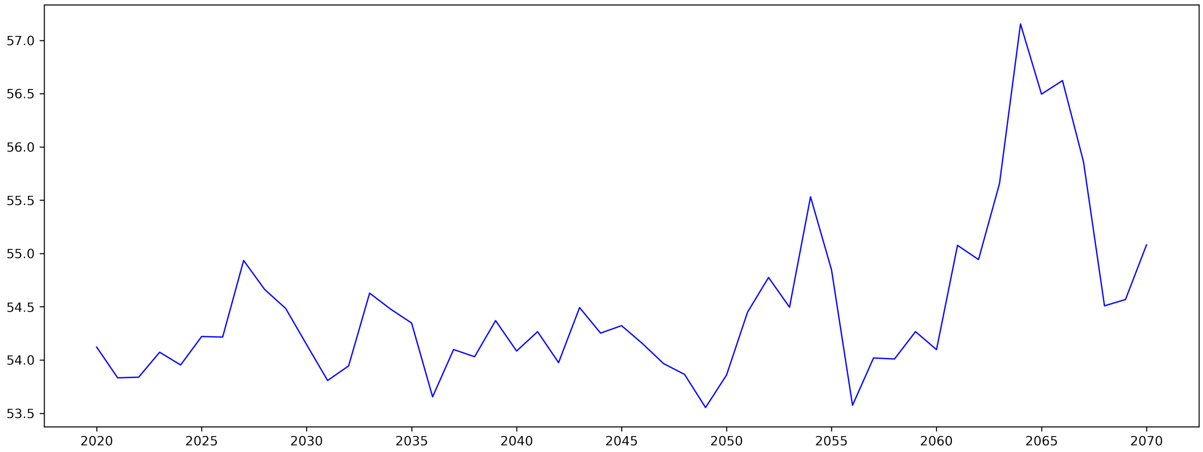
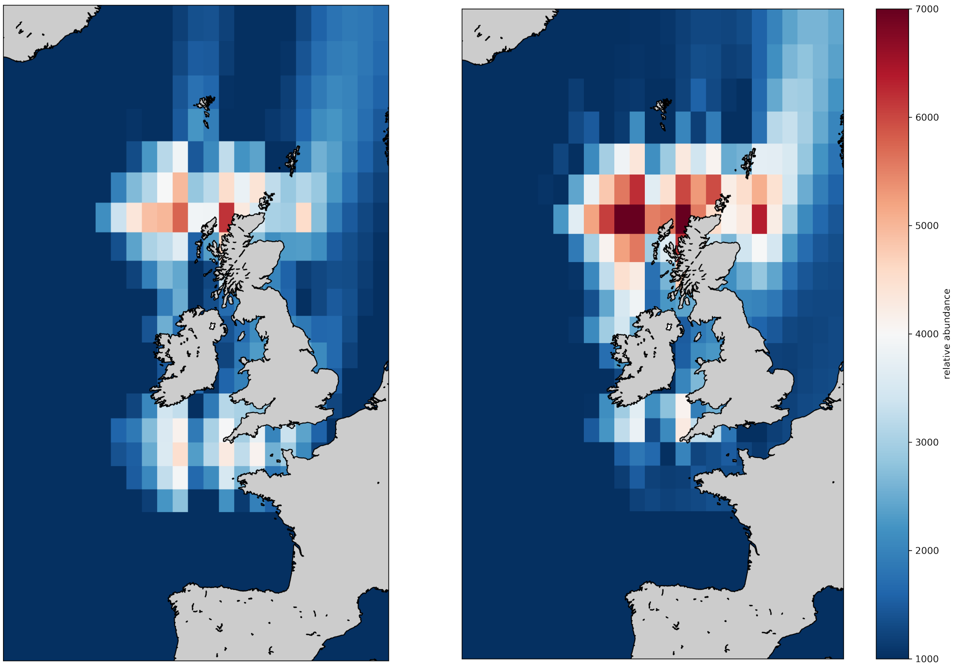


Figure 13: Latitudinal Shift in CoG of Herring for Problem 1

When we make projection about the future distribution of herring, things were interesting. As **Figure 14** indicates, with herring migrating toward the North, Scottish fishers will be able to catch more herring than before in the next 50 years.



(a) Herring Distribution in 2020 (b) Herring Distribution in 2065

Figure 14: Herring Distribution, Now and Future for Problem 1

### 4.2.2 Mackerel

When we predict future CoG of the mackerel as shown in **Figure 13** , what we got was a total mess. From the chart, we find the latitude CoG of Mackerel have a periodic change, which may have some relation ship with their food, the Plankton. The sea surface temperature is not the major factor of their habitat choice, but the temperature still has some effect on them, according to the chart. When the predicted SST gets its peak in around 2065, an evident shift is measured,

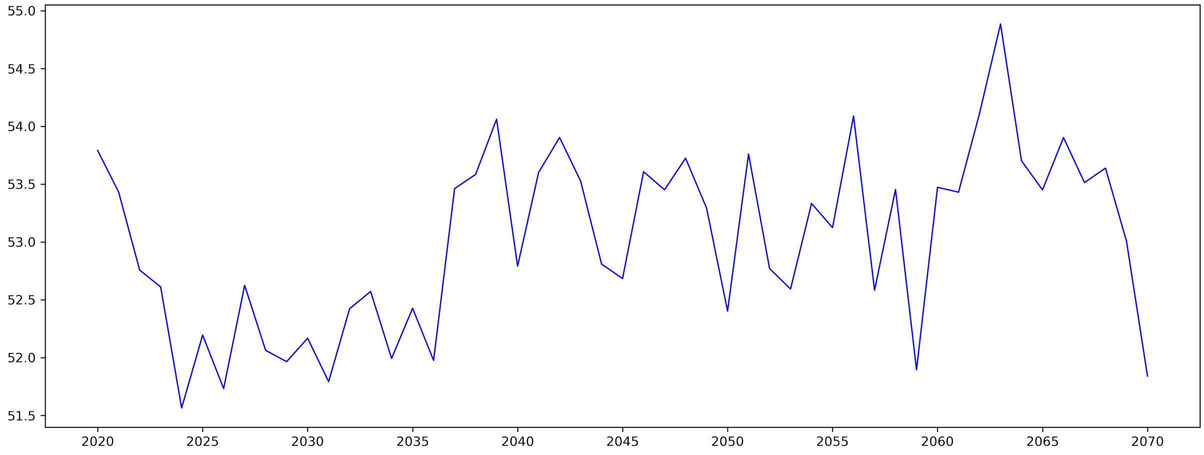
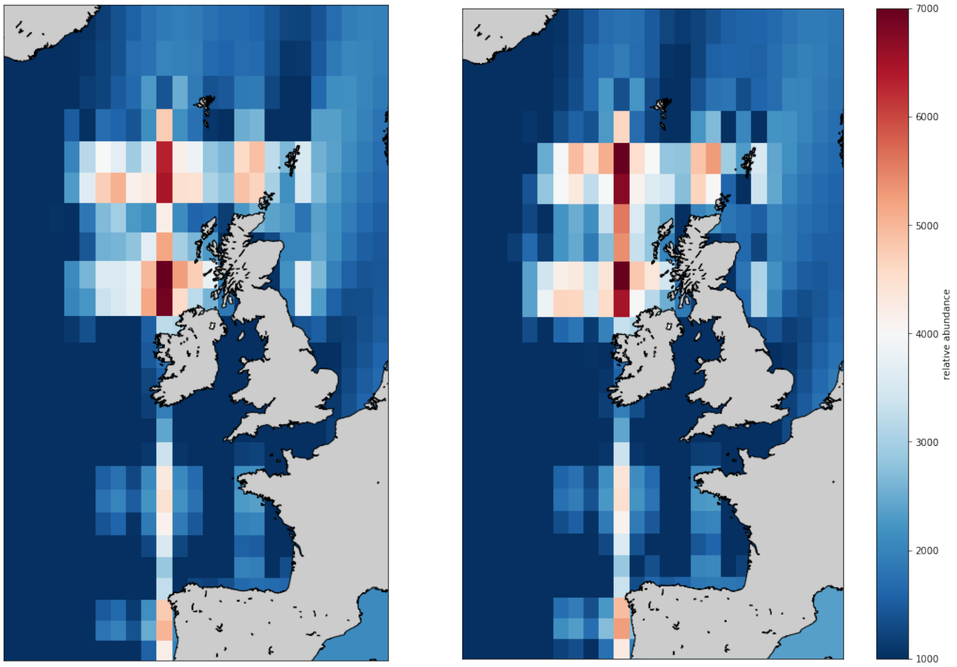


Figure 15: Latitudinal Shift in CoG of Mackerel for Problem 1

When it comes to the distribution, same conclusion is found. As shown in **Figure 16** . The Scottish can still get enormous benefits from the ocean in the next 50 years as before.



(a) Mackerel Distribution in 2020

(b) Mackerel Distribution in 2065

Figure 16: Mackerel Distribution, Now and Future for Problem 1

## 4.3 Prediction of the Time on Problem 2

By applying our model back to the historical data of the SST. We can build a error model $M_E$ of our model. The error data is also a time series data, hence we can fit the ARIMA model again on the error data to form the model.

Then, we can define the upper limit of the SST at a given grid as:

$$\overline{P(t)} = P(t) + M_E(t) \tag{19}$$

Where $P(t)$ is the result from the original model above, $M_E(t)$ is the result from the error model.

The lower bound as at a given confidence level $\alpha$ as:

$$\underline{P(t)} = \overline{P(t)} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1) \tag{20}$$

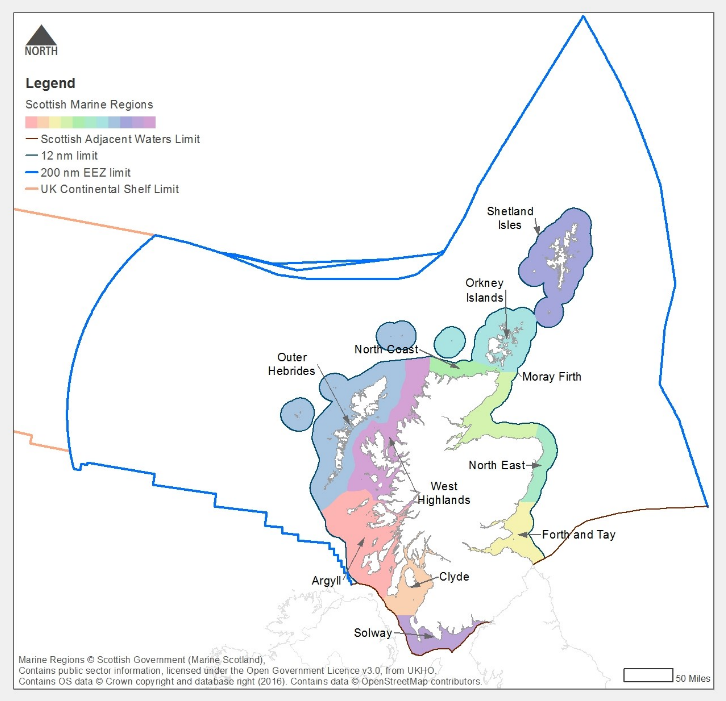
Using equation above with $\alpha$ as 95%, we can find the upper bound and lower bound of SST at the selected area. Put the new SST data in our fish shifting model, we can find the best case and worst case of the moving of the species. It will takes about 75.2 years at least, 94.7 years at most and most likely 86.2 years for all small fishing companies if they keep their location.

## 4.4 Strategies for Small Fishing Companies

According to our model, **no** changes are needed for those small fishing companies over the next 50 years. We find that both herring and mackerel are indeed moving to Northern regions, but the “speed” of migration is not as much as we thought. In fact the next several decades may be the “Golden Age” for fishers in Scotland with the marines originally living at Saint George’s Channel more Southern regions moving towards North Sea and Ireland Sea.

## 4.5 Territorial Water Problem

In our projections, theirs is no territory sea problem for the next 50 years if the EEZ of the Scotland shown in **Figure 17** is not changed. But nothing is consistent, our model may not reflect the real world. Even in our model, things are different in the next century. If some vital fishery resources moves in to the territorial sea of another country (Norway, in other word), our strategy should be changed to moving their company location to North or find a new abundant marine specie to harvest. Because that condition means that the population of the fishery moves too far for those fishers if no action is taken.



Marine Regions © Scottish Government (Marine Scotland), Contains public sector information, licensed under the Open Government Licence v3.0, from UKHO. Contains OS data © Crown copyright and database right (2016). Contains data © OpenStreetMap contributors.

Figure 17: The Scottish Exclusive Economic Zone

# 5 Merits and Limitations

## 5.1 Merits of Our Model

- **Data preprocessing.** Some data used in our model is missing, based on the conditions of missing data, proper transformations are taken to maintain the authenticity of data as much as possible.
- **Mode decomposition.** In our model, EEMD is introduced to decompose the historical environment data into multiple IMF series and then make predictions on each IMF to obtain a better result.
- **Credible prediction.** ARIMA time series model is used to make projections of SST. The ARIMA model has considered the data change of many factors compared with traditional models, making our results more credible.
- **Multiple factors considered.** When we build our model for predicting the future distributions of the marines, multiple factors including monthly sea surface temperature, the sea surface height, the sea surface salinity, the longitude and latitude of the grid are used. Making our prediction more accurate than models

considering temperature data only. And we don't let the data speak the truth, not by people's intuition such as the mackerel's best living temperature is a certain Celcius degree. Less priori can lead to the truth.

- **Integral statistics comparisons and analysis** Before we officially use GAM as our final model, we test other models which actually perform worse than GAM in statistics. In the meantime, we present the statistics indexes to support our research result.
- **Result visualization.** Driven by the genuine data, our results is well presented via a variety of charts.

## 5.2 Limitations of Our Model

- In our model, we may not make full use of all the data we found. For example, the historical data of the concentration of $CO_2$ is not used to predict the SST.
- The determination of some parameters and utility functions may be subjective. For example, the $p$s and $q$s of the ARIMA model.
- In our model, we take five factors to predict the fish density distribution. But in fact, there may be more ecological factors to genuinely effect the fish community's living environment. Anyway, to treat the nature nice should always be the correct answer for human being.

# Article

### Scottish Fishers May Face a Future of Losing Fish to Harvest, Study Finds

Many fishers from Scotland fishing herring and mackerel are projected to face the declining production over the next few decades —- unless they adapt to climate change by fishing in a different area or find a new different species to catch, according to a new unpublished research.

"We know that adapting climate change is hard for those small fishing communities, it will require some totally new approaches to fishing.", said by an anonymous researcher, "But change will become the new normal. "

Fishing is the economic lifeblood which is culturally important for many cities along the coast, in some cases for hundreds of years. But climate change is expected to have a huge impact on the distribution and habitat preference of marine species around Scotland, as the study points out.

The researchers studied how climate change will affect the distribution for two common marine species around Scotland. They used the most accurate model to predict how sea surface temperature and other major factors are likely to change. Then they use a most recent model to determine how those important commercial fish species are likely to move under the projected scenario. They also estimated whether the count in the ocean regions where the species are typically caught.

For both species studies, both of them are projected to migrating north by few hundreds kilometres by 2050–2060. That is to say, the count in the North regions are going to improve while others is deteriorating, according to the study.

While the species studied is shifting as the ocean getting warmer, fishers often limit on where they fishing based on their experience passed down orally. For example, they may choose where to fishing based on traditional fishing territories, type of vessel and gear and local or global fishery laws. But climate change is not in the list currently. In other words, it will be a hard–hit for traditional fishers in next few decades if no action were taken. For a small fishing company, total catches may drops by 15 percent at most over the study period if they keeps their historical dependence on species.

Fortunately, the fishers still have some countermeasures to take. They can shift the target of their fishing vessels to follow their target species, this may in turn affect their base location. If they are in fear of nostalgia, they can choose to focus on new abundant species in their traditional fishing regions. The study suggests.

Fish distribution changes will also have implications on fishery management for political reasons. Current example is the mackerel wars, which is political tensions between the EU, Norway and Iceland, created by the expansion of mackerel into Icelandic waters. As changes in the distribution of those commercial fish become evident in the future, cooperation with the neighbouring country to revise current fishery agreement is necessary, the study stresses.

# References

[1] Scottish Goverment. *Scottish sea fisheries statistics 2018*. Scottish Goverment, 2019.

[2] William B Monahan and Morgan W Tingley. Niche tracking and rapid establishment of distributional equilibrium in the house sparrow show potential responsiveness of species to climate change. *PloS one*, 7(7), 2012.

[3] ZHANG Ying, TAN Yan-chun, PENG Fa-ding, LIAO Xing-jie, and YU Yu-xin. Study on time series prediction model of sea surface temperature based on ensemble empirical mode decomposition and autoregresive integrated moving average. *Journal of Marine Sciences*, 2019.

[4] Wikipedia. Hilberthuang transform, 2020. [Online; accessed 16-February-2020].

[5] Wikipedia. Autoregressive integrated moving average, 2020. [Online; accessed 16-February-2020].

[6] Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(716-723), 1974.

[7] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Chapman & Hall/CRC*, 1990.

[8] Wikipedia contributors. Generalized additive model — Wikipedia, the free encyclopedia, 2019. [Online; accessed 17-February-2020].

[9] NOAA. Erddap. https://oceanwatch.pifsc.noaa.gov/erddap/index.html.

[10] Ices. http://www.ices.dk/Pages/default.aspx.

[11] Daniel Servén and Charlie Brummitt. pygam: Generalized additive models in python, March 2018.

[12] WJ Collins, N Bellouin, M Doutriaux-Boucher, N Gedney, T Hinton, CD Jones, S Liddicoat, G Martin, F OConnor, J Rae, et al. *Evaluation of the HadGEM2 model*. Met Office Exeter, UK, 2008.

# Appendix: Codes

## Prediction Pipeline in the First Problem

```
# In[1]:
#Herring Model is just of the same theory.
import pandas as pd
import numpy as np
from pygam import LinearGAM #the integral algorithm can be found in Github
import matplotlib.pyplot as plt

# In[2]:
# train a model:
mackerel_data=pd.read_csv("temp_mackerel.csv")
mackerel_data.head(10)

# In[3]:
X_train=mackerel_data.drop([’Value’,’Date’],axis=1).values
y_train=(mackerel_data[’Value’].to_numpy())
print(type(X_train))
print(type(y_train))
X_train.shape

# In[4]:
lams=np.random.rand(100,3)
lams=lams*3
lams=np.exp(lams)
gam=LinearGAM(n_splines=25).gridsearch(X_train,y_train,lam=lams)

# In[7]:
num=20
for num in range(20,71):
    target_file="future_test/rcp85-year/"+"20"+str(num)+".csv"
    dff=pd.read_csv(target_file)
    dff.dropna(inplace=True)
    dff[’Temp’]-=273.15

    X_test=dff.drop([’Date’],axis=1).as_matrix()
    y_test=gam.predict(X_test)
    dff[’predict_value’]=y_test
    dff=dff.loc[dff[’predict_value’]>0]

    dff.to_csv("diff_future_predict_mackerel/rcp85-year/"+"20"+str(num)+".csv")

# In[5]:
gam.summary()

# In[ ]:
titles = ["Lon","Lat","Temp_Diff"]
plt.figure()
fig, axs = plt.subplots(1,3,figsize=(40, 5))

for i, ax in enumerate(axs):
    XX = gam.generate_X_grid(term=i)
    ax.plot(XX[:, i], gam.partial_dependence(term=i, X=XX))
    ax.plot(XX[:, i], gam.partial_dependence(term=i, X=XX, width=.95)[1], c=’r’, ls=’--’
    if i == 0:
        ax.set_ylim(-30,30)
    ax.set_title(titles[i])
```

## Export history environment data from netCDF file

```
import netCDF4 as nc
import pandas as pd
import numpy

def date_inc(date):
    if date[1] + 1 == 13:
        date[0] += 1
        date[1] = 1
    else:
        date[1] += 1
    return date
```

```
file = '195912-200512.nc'

date = [1959, 12]
dataset = nc.Dataset(file)

lon1, lon2, lat1, lat2 = 340, 5, 40, 65


lat_dict = {}

for i, lat in enumerate(dataset['lat'][:]):
    lat_dict[int(lat)] = i

df = pd.DataFrame(columns=['Date', 'Lat', 'Lon', 'Temp'])

cnt = 0
for i in range(len(dataset['time'][:])):
    date_str = str(date[0]) + '-' + str(date[1])
    print(date_str)
    for lat in range(lat1, lat2 + 1):
        for lon in range(lon1, 360 + lon2 + 1):
            Lon = lon
            if Lon >= 360:
                Lon -= 360
            if type(dataset['tos'][i][lat_dict[lat]][Lon]) == numpy.float32:
                new = pd.DataFrame({
                    'Date': date_str,
                    'Lat': lat,
                    'Lon': Lon,
                    'Temp': dataset['tos'][i][lat_dict[lat]][Lon],
                }, index=[cnt])
                cnt += 1
            else:
                new = pd.DataFrame({
                    'Date': date_str,
                    'Lat': lat,
                    'Lon': Lon,
                    'Temp': None,
                }, index=[cnt])
                cnt += 1
            df = df.append(new, sort=False)

    date = date_inc(date)

df.to_csv('195912-200512.csv', index=False)
```