Time Series Analysis and Forecasting of Shellfish Contamination and Safety

André Pereira
andrespereira@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2021

Abstract

Harmful Algal Blooms (HABs) have been a rising issue not only due to environmental concerns, but also public health due to possible shellfish contamination. In Portugal, frequent analysis are ran by Instituto Português do Mar e Atmosfera (IPMA) to assess the quality of the water and its fauna, such as the shellfish and subsequently allow (or stop) its gathering and commercialization. These analyses, however, could be complemented and the swiftness of the fishing activity interdiction could be improved. For this, machine learning methods can be used to analyse temporal data (in the form of time series) in order to forecast the contamination of shellfish. This temporal data is gathered and compiled from the historical data present on IPMA's website which is released periodically at equal intervals, allowing a consistent time slices of the built time series. Several methods are presented and reviewed in this paper, which will be applied to collected data (that extend from the above mentioned time series to other environmental variables) in order to complement existing analysis work, which will also be extended through the usage of MAESTRO - an online tool for multivariate time series analysis. With this report and subsequent work - data collection, processing and forecasting, we will develop methods to support the prediction of shellfish contamination in Portugal's shoreline. No paragraph breaks.

Keywords: Harmful Algae, Marine Biotoxins, Time Series, Machine Learning; Forecasting

1. Introduction

Harmful Algal Blooms (HAB) are a worldwide concern becoming more frequent (and discovered, as some are still unknown and being found) and occurring in larger areas. Multiple poisoning syndromes exist and are derived from the consumption of shellfish contaminated with HABs - paralytic, diarrhetic, neurotoxic, amnesic and azaspiracid[23]. Most marine toxins are produced by dinoflagellates. An exception is the domic acid, the amnesic poisoning toxin, which is produced by diatoms of the *Pseudo-nitzschia* genus. [26]. Portugal's national monitoring of HAB's is done by the Portuguese Institute of the Sea and Atmosphere (IPMA - Instituto Português do Mar e Atmosfera) . The monitoring is done through various methods of biotoxin level surveys which lead to the different result bulletins published all over the world (complexity can even be different); these reports rely on a large amount observed data, from satellite imagery of ocean colour and historical trends to forecasts of bloom progression and even public health reports[26]. The portuguese HAB report is a weekly bulletin released in order to (in a concise and simple manner) inform on the harvestability of shellfish in the multiple zones of Portugal's coastline, which is divided in 9 main areas (L1-L9) as shown in Figure 1 some of which are subdivided into smaller areas.

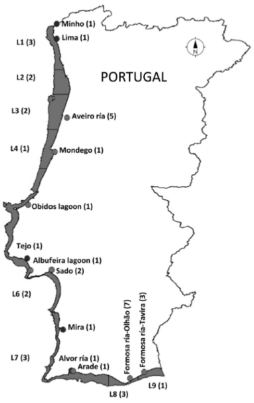


Figure 1: The 9 main areas of Portugal's coastline (with some respective subdivisions - totaling 40).

HAB rates have been increasing[9] and changing at alarming enough rates[11] to warrant more care and the development of more accurate studies in order to avoid the harvest of potentially contaminated shellfish and subsequently commercialize it, causing a public health issue that could have been avoided. There are a lot of sectors affected by this issue that are not obvious at first glance. Not only is this a complex concern that tackles many sectors and needs to be further researched by the scientific community over time, but a simple error in the analysis that deems a contaminated shellfish sample as marketable and consumable is a serious public health hazard that should be avoided at all costs[18][17]. These incorrect assessments do not end at a public health level but also on a production and market level - economic sectors, especially related to shellfish and marine food in general can have serious repercussions[22] and profit reductions.

Timing is essential and as such, early warning of HABs presence and its statistics - time, location (within the coastline areas) and magnitude is crucial information in order to control the coastal zones and the respective aquacultures and fishing practises in them; this allows to enhance business plan practises and ensures the best possible benefit for public welfare health wise[6]. Despite being a big concern, other factors must be collected and studied in order to accomplish the task of forecasting seafood contamination[15]. With this work, the collection of the necessary factors/variables and respective studying in the form of time series should provide the desired results in order to assist the various affected sectors (ranging from economical to public health) in the resolution of the issues mentioned above. The data was obtained from two key sources: IPMA's website itself which presents on a weekly basis a bulletin of the toxin levels in the shellfish in each area of the coast, the respective shellfish species and where the samples were taken. Copernicus is European Union's Earth observation programme; it studies the planet and its environment and offers information drawn from satellite observations and in-situ (non-space) data [1]. Copernicus will thus, be a valuable source of information to extract further data such as Chlorophyll and Sea Surface Temperature (SST).

2. Background

2.1. Time Series

Time Series (TS) are a series/collection of data points recorded through time in constant intervals, which are then modelled in order to determine patterns and the evolution of the series through time so as to forecast and predict future values[5]. A common notation to represent TS is the following:

$$X = \{X_t : t \in T\}, \tag{1}$$

where T is the index set.

Time series to be worked within this thesis will be both univariate and multivariate, with a focus on the former. Multivariate Time Series (MTS) consist of a time series where multiple variables change over time[14]. This differs from a Univariate Time Series where only one variable changes through time, as the name suggests.

2.2. Stationarity

A time series is stationary if its statistical properties (mean and variance) do not change in regular time intervals - there is no variable distribution over time. This is a property very useful for analyzing and modelling, so much that even most models assume this property in order to give a more complete analysis result.

2.3. Seasonality

Seasonality concerns certain patterns that occur frequently over time (called seasonal variation)[7]. Seasonality is important for the analysis of time-series because it can be removed or studied, the latter of which is preferable in this case, as it can give new (and more) information to improve the applied model's performance. In the case of this project, there are certain variables that can be grouped into certain seasonal clusters: temperature and moon phases (and consequently the tides of the sea), for example. Stationarity is correlated with seasonality in the sense that a seasonal time series is not stationary due to the seasonal aspect's presence causing the time series to change values at different times and thus, stripping it of its stationary property.

2.4. Autocorrelation Function (ACF)

Represents variability in the attributed by measuring and comparing observations with a lagged version of themselves and thus, determining pattern changes with the progression of time. It will be an important metric in this thesis to measure how accuracy measures should be applied to evaluate the quality of the models that will be reviewed further in this section. Autocorrelation is usually represented through a graphic to better help visualize how the time series works[13]. Eq.2 showcases how ACF is calculated, essentially being the result of the division between the covariance and variance for any lag of value k time steps preceeding time step $i \in T$.

$$r_k = \frac{\sum_{i=k+1}^{T}(y_i - y^{'})(y_{i-k} - y^{'})}{\sum_{i=1}^{T}(y_i - y^{'})^2} \tag{2}$$

2.5. Partial Autocorrelation Function (PACF)

Is similar to the ACF above, but partial autocorrelation only compares observations among time series variables and their lagged values without the correlation between all lags in between, so for example, the partial autocorrelation of a certain lag k is the equivalent of the autocorrelation between a variable y_i and the lagged value $y_i - k$ that does not have values for lags 1 through $k-1$ - those linear dependencies are not accounted for[13]. Eq.3 showcases this mathematically.

$$r_k = \frac{\sum_{i=k+1}^{T}(y_i - y^{'})(y_{i-k} - y^{'})}{\sum_{i=k+1}^{T}(y_i - y^{'})^2 \sum_{i=k+1}^{T}(y_{t-k} - y^{'})^2} \quad (3)$$

2.6. Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is one of the most commonly used information criteria - it is an estimator of model selection based on out-of-sample prediction error[30]. It focuses on selecting a model (out of the given set - a cadidate set) that minimises the relative amount of lost information; this criterion is defined by the following formula:

$$AIC = -2\ln(L) + 2p, \quad (4)$$

where L represents the likelihood under the evaluated model and p is the model's number of parameters.

2.7. Bayesian Information Criterion (BIC)

Another commonly used information criteria is the Bayesian Information Criterion - similar to AIC, it differs in the second component of its representation:

$$BIC = -2ln(L) + pln(n), \quad (5)$$

where L and p are, respectively, the same as the ones in AIC - the likelihood under the evaluated model and the number of parameters of the model[28]. BIC adds a new variable into account - n, which represents the sample size (number of instances of the train set the model is fitted for).

2.8. Mean Squared Error

The Mean Squared Error (MSE) is a loss function that measures the average of the squared difference between the forecast observations and the actual ones (the error). It is measured through the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2, \quad (6)$$

where x_i is the observed value, x_i' is the predicted value and n represents the length of the time-series. Due to it being a mean ($\frac{1}{n}\sum_{i=1}^{n}$) of the square of the error $(((x_i - x_i')^2))$, its aim is to select models that have the lower difference for each datapoint, thus a smaller MSE represents smaller average errors and thus, a better performing model.

2.9. Root Mean Squared Error

The Root Mean Squared Error (RMSE) is another metric that measures differences between sample values and their predicted versions by the trained model. It is written as:

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2]^{\frac{1}{2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2}. \quad (7)$$

2.10. Mean Absolute Percentage Error

The MAPE - Mean Absolute Percentage Error expresses the prediction accuracy of a model through the following ratio:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}(|\frac{y_i - y_i'}{y_i}|). \quad (8)$$

2.11. Autorregressive Model (AR)

An Autorregressive model (AR) is a regressive model that has its observations (values) depend on previous (lagged) observations - the variable is modeled through a linear combination of lagged values of that variable.

As such, an AR(p) model can be defined as:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + ... + \varphi_p x_{t-p} + a_t \quad (9)$$

Where $\varphi_1, \varphi_2, ..., \varphi_p$ stand for coefficient parameters, p stands for the number of lagged values used and a_t is the random term of the data (or white noise) which follows a white noise process (WN): $a_t \sim WN(0, \sigma^2)$. c represents a constant.

2.12. Moving Average Model (MA)

The MA model (or Moving Average Process) defines the output variable using a regression model on the past value errors - the lagged white noise values. An MA(q) model can be written as:

$$x_t = c + \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_p a_{t-p} + \varepsilon_t \quad (10)$$

Where q is the number of lagged values used (much like p for the AR model), $\theta_1, \theta_2, ...\theta_q$ are coefficient parameters, $a_t, a_{t-1}, ..., a_{t-q}$ are the white noise error terms[29]. Like the AR model, it can be re-written as:

$$x_t = c + \sum_{j=1}^{q}(\theta_j a_{t-j}) + a_t \quad (11)$$

2.13. Autorregressive Moving Average Model (ARMA)

The Autorregressive Moving Average (ARMA) model mixes both an AR(p) model and an MA(q) model and is thus usually written as ARMA(p,q). As it logically implies, it is a composition between the two previously mentioned and described models and can be expressed as:

$$x_t = \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{12}$$

Where δ is the constant term of the model, ϕ_i represents the autorregressive coefficient, θ_j is the moving average coefficient, ε_t illustrates the error term at time t and X_t is the observed value at time t [19].

Since ARMA is made of an AR(p) and MA(q) model combination, it is possible to generate its two counterparts due to the formula compositions:

- ARMA(p,0) is written as follows:

$$x_t = \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) + \sum_{j=1}^{0}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{13}$$

$$= \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) = AR(p) \tag{14}$$

- ARMA(0, q) leads to the following equation:

$$x_t = \delta + \sum_{i=1}^{0}(\phi_i x_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{15}$$

$$= \delta + \sum_{i=j}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t = MA(q) \tag{16}$$

2.14. Autor regressive Integrated Moving Average - ARIMA

ARIMA - Autorregressive (AR) Integrated (I) Moving Average (MA) model takes the core Autorregressive Moving Average model and combines both autorregressive and moving average processes building a model that also differences a time series in order to achieve its stationarity[19]. An ARIMA model is typically described as ARIMA(p,d,q) and is written as such:

$$(1 - \sum_{i=1}^{p}\alpha_i L^i)x_t = (1 + \sum_{i=1}^{q}\theta_i L^i)\epsilon_t \tag{17}$$

2.15. Random Forests

Random Forests (RF) are an ensemble learning algorithm[12] that build a set of decision trees that are then trained and are then used for classification or regression. By training each tree with a random set of data samples, the learnt results are the multiple uncorrelated trees built during training[2]. The trees will use a fixed value of features, randomly picked, to split the nodes and help with classification and/or prediction. By using a subset of the total features, combined with the usage of multiple trees, this minimizes the chances of overfitting (like a single tree would be more subject too) and thus, prediction errors associated with it[16]. The learnt smaller models (the trees) are then combined into a single prediction result. These methods can go from a majority voting for categorical attributes or an average for numerical attributes. In this work, Random Forests were used as regressors - each node splits into two other nodes until it reaches the leaves (the final node, determined by the RF's depth value - determined by the user), which have the average of the observations in them. The motivation to use RF's in this work was helped by existing research in this theme - Cheng et al.[4] used an Interative Random Forest (iRF) to determine the impact of nutrient conditions on algal abundance and also explore the interactions between microbial abundances and phytoplankton in order to better understand how bacteria and HABs interact with one another. The conclusions drawn proved inland nutrient fluxes were more relevant as the oceanic fluxes proved more volatile due to climate oscilations (and adding the variability of precipitation and upwelling). Other RF studies also proved fruitful: Valbi et al. [25] used an RF model to forecast paralytic toxin concentrations (*Alexandrium minutum*) in the Adratic Sea. By forecasting one week ahead of time and including upwards of 18 variables, the results were satisfying: the model correctly classified more than 85% cases of presence (or absence) of the (*Alexandrium minutum*) dinoflagellate. Furthermore, a second test was used where it lead to the conclusion that nutrient concentrations are not needed to ensure an a high-performing model so the second model was preferred during the study for practical issues.

2.16. Bayesian Networks (BN)

BN's are statistical models that represent attributes and their conditional dependencies in a directed acyclical graph (DAG)[27]. Bayesian Networks are very effective in the prediction of the likelihood of specific attributes triggering a certain outcome in an event as they use Bayesian inference to model conditional dependence through edges that connect the related variables through nodes thus creating a DAG that models causation between the variables of a dataset.

Taking the example with node C, which splits into child nodes D and E, we can see an edge connecting C and D so, $P(D|C)$ is a probability to be taken into account in joint probability distributions - this way, probabilities associated with B and A (C's parent nodes) must be known to calculate any inferences related to these attributes.

To give a simple example, we can write the Bayes rule of posterior probability, $P(D|C)$, given P(D) and the likelihood $P(C|D)$:

$$P(D|C) = \frac{P(C|D)P(D)}{P(C)} \tag{18}$$

Which can be simplified into one of the fundamental rules of probability:

$$P(D|C) = \frac{P(D,C)}{P(C)} \tag{19}$$

In the case of conditional independence, such as D and E, then we can simplify certain probabilities that involve these conditionally independent variables:

$$P(D|C,E) = P(D|C). \tag{20}$$

This way, we can define the whole structure of a BN by specifying the probability distributions of all nodes with parents and the probabilities of the root node (or nodes, should there be more than one).

2.17. Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are a generalization of Hidden Markov Models which can be represented as the simplest form of a DBN [21]. Due to the time properties of the data of this work, Bayesian Networks do not work very well in representing these temporal dependencies that are so characteristic of time series; they are, however, a good base for Dynamic Bayesian Networks which can actually model and work with data that is time dependent (that evolves over time and can be called dynamic as a result of that, thus the name).

DBNs extend the regular BN notion to allow modelling of time influences, ergo, modelling dynamic systems/data such as the time series in this work. Similar to a BN, comprised on nodes and edges, the DBN formally introduces time slices into the network's architecture, as now there's a temporal connection between variables and thus, conditional probabilities exist between variables at different time slice points. it is worth noting DBNs follow the first order Markov property of only the immediate past affects the state of a system at any time slice *t*. So, for any node *x* in the network's node set, a transition from time slice *t-1* to *t* has the probability

$$P(x_t|x_{t-1}) \tag{21}$$

for any node *x* in the network's node set. It is, intuitively, well suited to represent markov processes. So we can represent the join distribution through a chain of time slices for a certain variable X:

$$P(X_{1:T}) = P(X_T|X_{1:T-1}) \tag{22}$$

Where $X_{1:T}$ is a sequence X_1, ..., X_t, ..., X_T

Overall, a DBN is a factorisation of a probability distribution where time slices are present, through composite states at each time slice *t*. Variables in different time slices can have relations between them, thus originating more edges in the network. A DBN factorisation can be written as:

$$P(X_{1:T}) = \prod_t \prod_i P(X_{t,i}|pa(X_{t,i})) \tag{23}$$

Where i groups variables in a same time slice and pa(X) represents the parents of X in the network. In the field of medical data analysis, it is frequent to adopt the simpler first order Markov property in order to simplify the model, making the future dependent only on the present. Intuitively, it makes sense as the present health status gives the better information about the future status, and not the past ones. A similar approach should be used for the context of the data analysis required in this thesis.

2.18. Artificial Neural Networks

ANN implementations were originally aimed at solving problems in a similar manner to the human brain but over time they have proven excellent in certain fields, such as biology and speech recognition. Neurons are represented by nodes and are connected between edges. Neurons and edges usually have weights assigned to them that are adjusted the further the NN is trained. Neurons are aggregated into layers and thus, ANN's typically have three main layers: the input layer, the hidden layer and the output layer.

The input received in the first layer is processed in the hidden layer which is made of several neurons, possibly spread between several sub-layers; each value is affected by an activation function present in the hidden layer's neurons, which can be a sigmoid for example, among many others; different layers can have different input transformations and are then sent through the connected edge to the next set of neurons to repeat the process.

There have been researches done in this field, namely in Recknagel et al. (1997) where ANNs were trained to forecast and try to prevent or detect in time an algal bloom[20] in lakes. One of the lakes, Lake Kasumigaura, obtained great results, having its ANN predict the timing, magnitude and succession of algal blooms, even using independent data not used in the training process.

2.19. Gradient Boosting (and XGBoost)

Gradient Boosting is a technique that works for both classification and regression alike - it essentially ensembles weak prediction models (such as decisions trees, which will be used here) into stronger ones by optimizing the model performance[8]. The ensemble part is similar to the one seen in a

Random Forest Regressor already approached - it builds a final model from the combination of learnt smaller/individual models.

The gradient component derives from the typical Gradient Descent seen in Neural Networks - multiple model predictions are combined in order to iterate improvements on following assembled trees.

Chen, et al.(2016)[3] studied Friedman's Gradient Boosting documentation [8] and developed XGBoost, achieving a state-of-the-art machine learning method that has proven vastly effective in both regression and classification supervised problems.

Describing their algorithm, it uses K additive functions are used to predict an output through a tree ensemble model:

$$y'_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{24}$$

where F is the space of regression trees (CART). XGBoost proceeds to learn the functions used by minimizing a regularized function:

$$L = \sum_i l(y'_i, y_i) + \sum_k \Omega(f_k), \tag{25}$$

where l is a differentiable convex loss function that measures the difference between the prediction and the actual value (y'_i and y_i respectively) - this is the case because it's easier to use a convex loss function to find global optimums (since we're speaking of loss functions, these optimums are generally represented as minimums). A property of these functions is that local minimums are global minimums thus optimization algorithms like the gradients used here, can be used to find optimal results globally. Ω is the model complexity that serves to regularize trees. It is defined as:

$$\Omega(f) = \gamma T + \frac{\lambda w^2}{2}, \tag{26}$$

Here, γ represents a gain threshold - should the calculated gain surpass γ's value, then that branch can be generated (partition of a leaf node) as it has sufficient gain. λ portrays a regularization parameter (L2) and helps to avoid over-fitting.

XGBoost has shown excellent performance in both forecasting, such as with crude oil prices(Gumus et al. [10]) and classification, such as Torlay et al.[24] which managed to classify patients with epilepsy with an AUC (Area-Under-Curve) mean score of 91%.

3. Implementation

The process of data collection proved to be a bigger challenge than expected as, for instance, the time series changed over time for numerous reasons (this following list will refer to values seen in the Lipophilic Toxin values of the biotoxins as they were the ones mainly studied in this work):

Zones changed names over the years and some were even introduced throughout the years while toxin value thresholds also changed over time - earlier in 2015/2016 being 850 (any value above it was simply referenced as 850 μg per kilogram of okadaic acid and equivalent toxins. In 2017 that value was lowered to 625 and since 2018 it has stayed at an even lower value of 550 (with no changes regarding units and measures). These value changes do not affect the study of the series in a major manner but it is worth noting that valuesof previous thresholds, such as 625 or even 800 would be accounted for originally while now, those values will be regarded as the much lower value of 550, which can affect some model performance due to the reduced range of values, leading to possible missing value fluctuations that we could observe in 2015's threshold value of 850.

Regarding the phytoplankton data, other challenges needed to be taken into account, such as:

Other species of Phytoplankton have started being accounted for and quantified in the monthly IPMA report. Inititally (in 2014) the reports approached quantifications of specific species which were replaced since late 2014 by a generalization - DSP producing phytoplankton were all bundled into a single variable (no information on which species were studied are present) and the same applied to ASP and PSP producing species. In early 2017, 2 new categories were added: Yessotoxin and Azaspiracid producing species. Starting 2018, the Azaspiracid category was removed and later in May 2018 was added back, alongside 5 new variables. The existing variables were altered and split as the monthly data changed into 10 total variables that now mention the class of phytoplankton and the respective toxins they produce.

Also starting in May 2018, data values also changed. Before, values were frequently marked as zero in the tabular data, signifying that an area has no toxin-producing algae of that category. After May 2018, however, data became frequently marked as $< LD$ which means *Below Threshold* , replacing the zeroes seen in the data until that point.

Biotoxin data also has values that are categorical instead of numerical and had to be replaced in the data; these values are:

- ND represents a value that is deemed *Not-Detected* as the analysis devices couldn't detect the little to no amount present in the collected shellfish sample.

- NQ dictates the analysis sample has a toxin rate that was detected but was too low to be quantified (thus, NQ stands for Not-Quantifiable).

- NR is the final categorical value that means Not-Done, meaning the sample wasn't analysed and as such, this logically represents a missing value in our data.

For the PDF data available online, an automatic extraction tool was developed to allow the user, through a simple command line, to extract any data file directly from the IPMA website and save it locally - an added feature of conversion from PDF to CSV format is also present to not only download the data, but download it in a format easier to process.

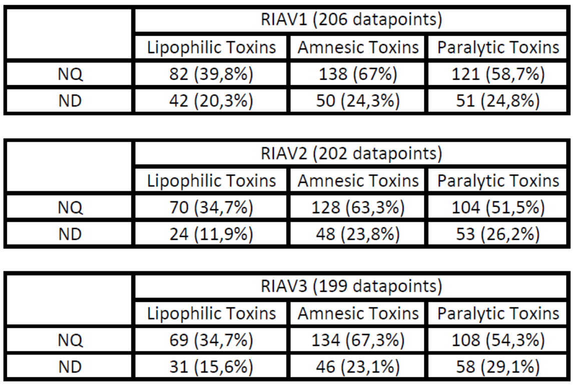
All the above information was related to the data collection and treatment process of IPMA's data files but more work was dedicated to acquiring further data related to missing (but possibly meaningful and correlated) features and as such, throughout development Copernicus data was also gathered. The features extracted were the chlorophyll and Sea Surface Temperature (SST) values which were then appended to the various time series used in this study.

For pre-processing, the first thing required was filling the missing values as mentioned above and the following graphics and analysis will have its missing values replaced by the mean of the remaining values in the dataset.

One of the first things was to see how the toxin rates evolved over time. Additionally, more attention was given to the Lipophilic toxins as they are the most predominant in Portugal and suffered the most changes over time - Amnesic and Paralytic had very few noticeable variations over the course of the 4 year dataset that was studied, proving to be much less fruitful datasets to work with.

The Wedge-Clam L8 dataset was chosen to be studied and forecast. It suffered a train/test split on the start of November 2019 - November and December consisted of seven total data points the models would be evaluated on - the remaining time period before that is the training set and is comprised of the remaining 159 data points - forecasting results will be shown in the Results section.

For a multivariate time series analysis process, an online Time Series Analysis through Dynamic Bayesian Networks was used: MAESTRO. The same pre-processing procedures seen in Chapter 3 were applied to the RIAV dataset for the cockle shellfish - specifically, RIAV1, RIAV2 and RIAV3. These were also complemented using SST, Chlorophyll and their respective phytoplankton data, making the dataset use a total of eight variables to observe how MAESTRO's modeled trees linked these variables among each other (and obtain possible causal relations between them).



	RIAV1 (206 datapoints)		
	Lipophilic Toxins	Amnesic Toxins	Paralytic Toxins
NQ	82 (39,8%)	138 (67%)	121 (58,7%)
ND	42 (20,3%)	50 (24,3%)	51 (24,8%)

	RIAV2 (202 datapoints)		
	Lipophilic Toxins	Amnesic Toxins	Paralytic Toxins
NQ	70 (34,7%)	128 (63,3%)	104 (51,5%)
ND	24 (11,9%)	48 (23,8%)	53 (26,2%)

	RIAV3 (199 datapoints)		
	Lipophilic Toxins	Amnesic Toxins	Paralytic Toxins
NQ	69 (34,7%)	134 (67,3%)	108 (54,3%)
ND	31 (15,6%)	46 (23,1%)	58 (29,1%)

Figure 2: Below Threshold (ND and NQ) value counts in the RIAV1, RIAV2 and RIAV3 time series, with respective rate percentage.

4. Results

4.1. Forecasting for the L8 zone

For the forecasting of the L8 Wedge-Clam time series, several models were applied. Figure 3 shows the model's forecasting of the test set.

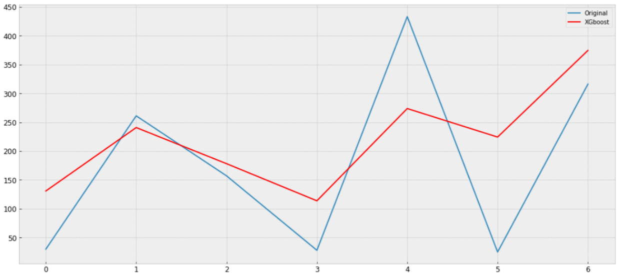


Figure 3: XGBoost forecasting performance on the time series

Besides XGBoost, a Random Forest Regressor was used, as well as an ARIMA model.

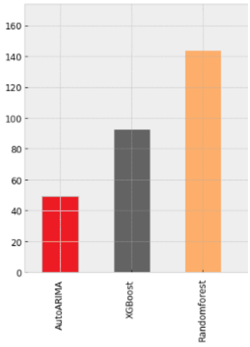


Figure 4: MAE metric results for each model trained with the Wedge Clam dataset on the L8 area.

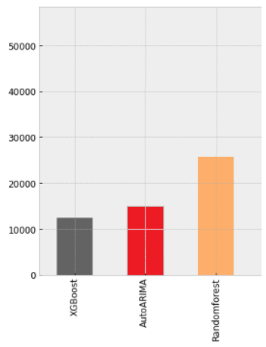


Figure 5: MSE metric results for each model trained with the Wedge Clam dataset on the L8 area.

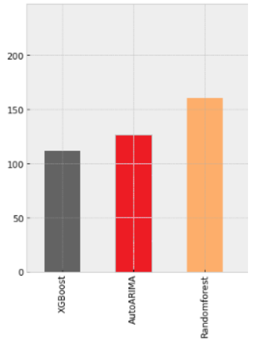


Figure 6: RMSE metric results for each model trained with the Wedge Clam dataset on the L8 area.

Figure 4 shows that the Mean Absolute Error (MAE) was relatively low in 2 particular models, who achieved a score of under 100, those being the XGBoost and AutoARIMA model. As for the MSE, Figure 5 shows that AutoARIMA and XGBoost models obtained a respectable value that complements their good performance on the MAE evaluation and the RMSE evaluation, seen in Figure 6.

4.2. Multivariate time series analysis using MAESTRO

For this analysis the RIAV dataset was used, specifically, RIAV1, RIAV2 and RIAV3, using the cockle shellfish as the species to study due to its high amount of samples. The reason RIAV 4 was not taken into account was because of its considerably lower number of samples compared to the previous three, which have very similar amounts of data points (206, 202 and 199, in order). RIAV4's inclusion in this analysis would prove very complicated due to the sample size disparity and would likely cause more error-prone results.

The resulting DBN models and condition probability tables of the three RIAV time series show that because of the above mentioned high lack of recorded values outside of the detection (or quantification) threshold, there's a very big similarity among the phytoplankton concentrations and the resulting model approaches those relations as they are naturally far stronger than other attributes (such as temperature, chlorophyll or even lipophilic toxin concentrations in cockles) that have a higher rate of recorded values outside any lower (or higher) thresholds. However, some interesting inferences were detected in the models - in the RIAV2 time series, the temperature (SST) seemingly the temperature seemingly has an influence on the PSP producing phytoplankton's concentrations when both are in a higher bin and sea surface temperatures starts decreasing. The RIAV3 time series showed that the sea surface temperature seemed another factor that influenced the resulting amnesic toxin concentrations when paired with the chlorophyll rates. The lower the SST (for the same chlorophyll values), amnesic toxin probabilities point to lower concentration values - this is especially noticeable when the lagged data point of the temperature is 0, meaning the recorded temperature at the time of the collected sample (toxin or phytoplankton).

After this study of the 3 different RIAV zones, the logical next step was to evaluate any possible correlations between the data present in two zones, so the datasets needed to be combined. For this purpose, RIAV2 and RIAV3 time series were combined using a familiar process done before: because MAESTRO requires multiple time series to be together, the dates needed to be processed in order to allow the joining process of the zone time series and thus, time series datapoints were joined on the closest date that did not exceed a set threshold of a week. Since RIAV2 and RIAV3 data were usually collected in the same week, likely due to their geographical proximity, this method made the most sense.

While a big portion of the results obtained from the DBN model generated (see Figure 7) and the respective conditional probability tables yields inconclusive relations, there is a considerable correlation between RIAV2 and RIAV3's chlorophyll rates. Higher values verified in RIAV3's chlorophyll quantities seem to yield higher values in RIAV2's chlorophyll values - the exact same applies for lower values. Given this information, paired with the other analysis performed on the single RIAV time series (in particular RIAV2 and RIAV3), we can observe a probable correlation between chlorophyll, sea surface temperature and biotoxin or phytoplankton concentrations, as seen in RIAV2 with the PSP producing phytoplankton and sea surface temperature and in RIAV3 with the pairing of chlorophyll and SST regarding amnesic toxin concentrations found in the cockle samples.

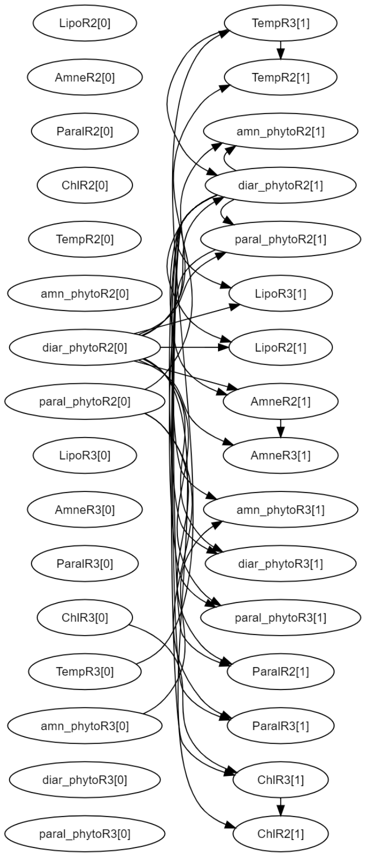


Figure 7: MAESTRO's resulting DBN model for the joined time series of RIAV 2 and RIAV3

5. Conclusion and Future Work

5.1. Conclusion

IPMA's analysis serves as the frontline to prevent the harvesting and subsequent commercialization (and consumption) of contaminated shellfish. This is done through the analysis of shellfish samples (for biotoxin concentrations in them) and HAB presence in collected water samples - should these analysis results go over the legal limit, the affected zones (of which there are fourty across the entire portuguese coast) are shut down temporarily until another sampling proves the contamination is no more. With this work, the aim was to enhance the swiftness of the zone blocking through methods of forecasting in order to determine zones that could have contaminated shellfish ahead of time. A brief revision of some methods applied in this thesis were studied - including some related work where they were used and proved to be effective. Furthermore, a brief examination of concepts related to time series were presented in order to better understand the thought process in the developed set of models and data processing. Through the development of forecasting models and with the assistance of MAESTRO, a better understanding of the shellfish contamination and its causes were achieved - namely a correlation that indicates sea surface temperature and chlorophyll had an influence in the amnesic toxins found in cockles in the RIAV3 zone and in the PSP producing phytoplankton in the RIAV2 zone; when joining two time series from different zones (RIAV2 and RIAV3 specifically), chlorophyll values from RIAV3 seemed to directly correlate with the values seen in RIAV2. With the above described analysis and the pre-processing and collection of the data provided by IPMA, it is hoped that the accessibility for further work in this field can be done in order to enhance the analysis already done here and further reach the optimal goal of consistently (and accurately) predicting biotoxin contamination in shellfish, no matter the species or the region the sampling was done.

5.2. Future Work

With the data collected and processed, there are time series with very few data points which make accurate predictions far harder. For this, the development of models optimized for these smaller time series would extend this forecasting work for more regions and species and thus, cover more potential contamination events. Still pertaining the model suggestions, more models could be developed to test their performance in these datasets, such as Long-Short-Term-Memory Neural Networks or Gaussian Process Regression (or Kriging). Likewise, showcasing these time series, paired with the respective forecast models in a possible web application would prove fruitful for both the easiness of studying these time series, but also for accessibility to the workers of possible affected sectors (such as fishing and commerce), and even the civilian population. More attributes could also have been added - salinity, currents and rainfall are examples of possible factors that could affect the forecasting results. An extended collection of attributes to add to the existing time series could add valuable correlations between biotoxin contaminations, HABs and the various factors that affect the coastal areas and their dynamic.

References

[1] Copernicus - marine environment monitoring service, 2020. Online; 26 December 2020.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[4] Y. Cheng, V. N. Bhoot, K. Kumbier, M. P. Sison-Mangus, J. B. Brown, R. Kudela, and M. E. Newcomer. A novel random forest approach to revealing interactions and controls on chlorophyll concentration and bacterial communities during coastal phytoplankton blooms. *Scientific reports*, 11(1):1–11, 2021.

[5] M. Dastorani, M. Mirzavand, M. T. Dastorani, and S. J. Sadatinejad. Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition. *Natural Hazards*, 81(3):1811–1827, 2016.

[6] K. Davidson, D. M. Anderson, M. Mateus, B. Reguera, J. Silke, M. Sourisseau, and J. Maguire. Forecasting the risk of harmful algal blooms, 2016.

[7] P. H. Franses. Seasonality, non-stationarity and the forecasting of monthly time series. *International Journal of forecasting*, 7(2):199–208, 1991.

[8] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[9] C. J. Gobler, O. M. Doherty, T. K. Hattenrath-Lehmann, A. W. Griffith, Y. Kang, and R. W. Litaker. Ocean warming since 1982 has expanded the niche of toxic algal blooms in the north atlantic and north pacific oceans. *Proceedings of the National Academy of Sciences*, 114(19):4975–4980, 2017.

[10] M. Gumus and M. S. Kiran. Crude oil price forecasting using xgboost. In *2017 International conference on computer science and engineering (UBMK)*, pages 1100–1103. IEEE, 2017.

[11] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull. Habnet: Machine learning, remote sensing based detection and prediction of harmful algal blooms. *arXiv preprint arXiv:1912.02305*, 2019.

[12] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[13] A. Inoue. Asymptotic behavior for partial autocorrelation functions of fractional arima processes. *Annals of Applied Probability*, pages 1471–1491, 2002.

[14] S. S. Jones, R. S. Evans, T. L. Allen, A. Thomas, P. J. Haug, S. J. Welch, and G. L. Snow. A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics*, 42(1):123–139, 2009.

[15] S. Lee and D. Lee. Improved prediction of harmful algal blooms in four major south korea's rivers using deep learning models. *International journal of environmental research and public health*, 15(7):1322, 2018.

[16] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[17] M. Mateus, G. Riflet, P. Chambel, L. Fernandes, R. Fernandes, M. Juliano, F. Campuzano, H. De Pablo, and R. Neves. An operational model for the west iberian coast: products and services. *Ocean Science*, 8(4), 2012.

[18] J. Nicolas, R. L. Hoogenboom, P. J. Hendriksen, M. Bodero, T. F. Bovee, I. M. Rietjens, and A. Gerssen. Marine biotoxins and associated outbreaks following seafood consumption: Prevention and surveillance in the 21st century. *Global food security*, 15:11–21, 2017.

[19] D. R. Osborn and J. P. Smith. The performance of periodic autoregressive models in forecasting seasonal uk consumption. *Journal of Business & Economic Statistics*, 7(1):117–127, 1989.

[20] F. Recknagel, M. French, P. Harkonen, and K.-I. Yabunaka. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*, 96(1-3):11–28, 1997.

[21] S. J. Russell and P. Norvig. Artificial intelligence: a modern approach. malaysia, 2016.

[22] I. Sanseverino, D. Conduto, L. Pozzoli, S. Dobricic, T. Lettieri, et al. Algal bloom and its economic impact. *European Commission, Joint Research Centre Institute for Environment and Sustainability*, 2016.

[23] A. Silva, L. Pinto, S. Rodrigues, H. De Pablo, M. Santos, T. Moita, and M. Mateus. A hab warning system for shellfish harvesting in portugal. *Harmful Algae*, 53:33–39, 2016.

[24] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu. Machine learning–xgboost analysis of language networks to classify patients with epilepsy. *Brain informatics*, 4(3):159–169, 2017.

[25] E. Valbi, F. Ricci, S. Capellacci, S. Casabianca, M. Scardi, and A. Penna. A model predicting the psp toxic dinoflagellate alexandrium minutum occurrence in the coastal waters of the nw adriatic sea. *Scientific reports*, 9(1):1–9, 2019.

[26] P. Vale, M. J. Botelho, S. M. Rodrigues, S. S. Gomes, and M. A. d. M. Sampayo. Two decades of marine biotoxin monitoring in bivalves from portugal (1986–2006): a review of exposure assessment. *Harmful Algae*, 7(1):11–25, 2008.

[27] M. Van der Heijden, M. Velikova, and P. J. Lucas. Learning bayesian networks for clinical time series analysis. *Journal of biomedical informatics*, 48:94–105, 2014.

[28] S. I. Vrieze. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, 17(2):228, 2012.

[29] D. B. Woodard, D. S. Matteson, S. G. Henderson, et al. Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics*, 5:800–828, 2011.

[30] Y. Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.