TÉCNICO LISBOA

Time Series Analysis and Forecasting of Shellfish Contamination and Safety

André da Silva Pereira

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Susana de Almeida Mendes Vinga Martins
Prof. Marta Isabel Belchior Lopes

Examination Committee

Chairperson: Prof. Luís Manuel Antunes Veiga
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Bruno Emanuel Da Graça Martins

October 2021

Acknowledgments

I would like to thank my supervisors, Professor Susana Vinga and Marta Lopes for their infinite patience and support when I faced my hardest academic challenges during the development of this work.

I must thank all my friends, even those who simply asked how my work was going - it meant a lot during the most challenging phases of this thesis. Likewise, a big thank you to my work colleagues for their equally remarkable support.

Last but not least, I'd like to thank my parents, who were available for me whenever I needed, no matter the hours - the continuous support from them and everyone deserves much more than this simple heartfelt acknowledgement.

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.

Keywords

Harmful Algae; Marine Biotoxins; Public Health; Time Series; Machine Learning; Forecasting; Analysis.

Resumo

A proliferação de algas nocivas tem sido um problema crescente, não só devido a consequências ambientais, mas também de saúde pública devido à possíveis contaminações de marisco. Em Portugal, análises frequentes são efetuadas pelo Instituto Português do Mar e Atmosfera (IPMA) para avaliar a qualidade da água e da sua fauna, tal como o marisco e subsequentemente permitir (ou impedir) a sua colheita e comercialização. Isto dito, estas análises podiam ser complementadas e consequentemente, a celeridade da interdição da atividade de pesca podia ser melhorada. Para isto, métodos de aprendizagem automática podem ser usados para analisar os dados temporais (sob a forma de séries temporais) para poder prever a contaminação do marisco. Estes dados são obtidos e compilados a partir dos dados históricos presentes no website do IPMA, que lança estes dados periodicamente em intervalos iguais, permitindo que os dados temporais da série sejam equidistantes entre si. Vários métodos são aqui apresentados e estudados neste documento, que serão seguidamente aplicados aos dados obtidos de forma a complementar o material de análise existente, que também será fundamentado com o uso do MAESTRO - uma ferramenta online para análise de séries temporais multivariadas através de redes bayesianas dinâmicas. Com este documento e o trabalho incluído - obtenção de dados e respetivo processamento, análise e previsão, os métodos desenvolvidos deverão suportar a previsão de contaminação tóxica de marisco na costa portuguesa de forma a permitir a proteção do comércio e assegurar a saúde pública relativamente ao consumo de marisco.

Palavras Chave

Algas Nocivas; Biotoxinas Marinhas; Saúde Pública; Séries Temporais; Aprendizagem Automática; Previsão; Análise.

Contents

List of Figures

List of Tables

Acronyms

HAB	Harmful Algal Bloom
IPMA	Instituto Português do Mar e Atmosfera
DSP	Diarrhetic Shellfish Poisoning
ASP	Amnesic Shellfish Poisoning
PSP	Paralytic Shellfish Poisoning
SST	Sea Surface Temperature
TS	Time Series
MTS	Multivariate Time Series
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
MSE	Mean Squared Error
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Squared Error
ACF	Autocorrelation Function
PACF	partial Autocorrelation Function
ADF	Augmented Dickey-Fuller
MA	Moving Average
ARMA	Autorregressive Moving Average
ARIMA	Autorregressive Integrated Moving Average

AR	Autorregressive
RF	Random Forests
iRF	Iterative Random Forests
GI	Gini Importance
BN	Bayesian Network
DBN	Dynamic Bayesian Network
DAG	Directed Acyclic Graph
NN	Neural Network
ANN	Artificial Neural Network
CSV	Comma Separated Values
NQ	Non-Quantifiable
ND	Non-Detectable
NR	Not-Done
MAESTRO	Dynamic Bayesian Networks Online.

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 [4]. 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. [5]. The main species reported by the Portuguese monitoring program of HABs belong to the genera *Pseudo-nitzschia, Dinophysis, Gymnodinium* and more recently *Ostreopsis* and *Karenia* - all these have different oceanographic preferences that then allow them to display different life-form (morphotype) characteristics and adaptive strategies [5]. 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 [5]. 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 1.1 some of which are subdivided into smaller areas.

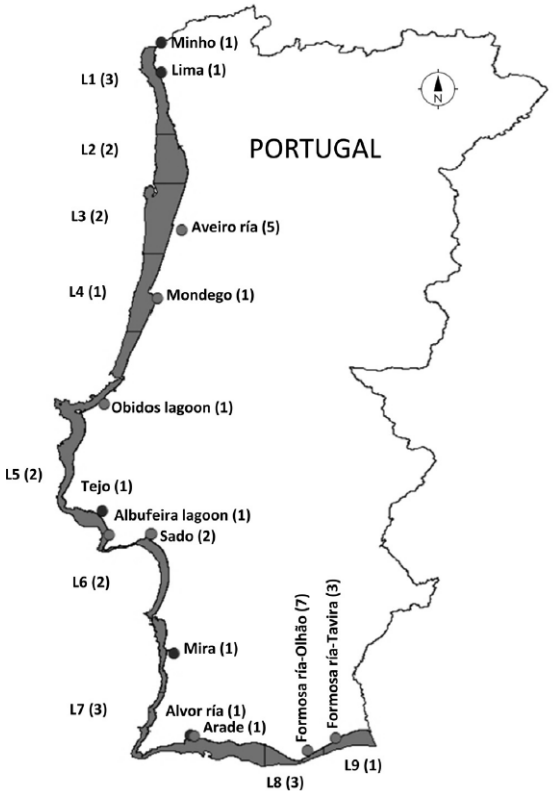

Figure 1.1: The 9 main areas of Portugal's coastline (with some respective subdivisions - totaling 40).

1.1 Motivation

Despite only 0.02% of phytoplankton species are capable of producing marine biotoxins, these toxins are a serious public health hazard. Moreover, due to global warming and general weather condition changes along the years, HAB rates have been increasing [6] and changing at alarming enough rates [7] [8] 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 [9] [10]. These incorrect assessments do not end at a public health level but also on a production

Toxins	Legal Threshold	Legal Reference
Amnesiac Toxins (**ASP**)	20 mg of domoic acid / kg	A
Lipophilic Toxins (Okadaic Acid, Dinophysistoxins and Pectenotoxins) (**DSP**)	160 µg of Okadaic Acid equivalent / kg	A
Lipophilic Toxins (Yessotoxins) (**DSP**)	3.75 mg of yessotoxin equivalent / kg	F
Lipophilic Toxins (Azaspiracid) (**DSP**)	160 µg of azaspiracid equivalent / kg	A
Paralyzing Toxins (**PSP**)	800µg/kg	A

Figure 1.2: Main hazardous biotoxins in the portuguese coastline and their respective maximum legal rates per kilogram of shellfish. Adapted and translated from IPMA [1].

and market level - economic sectors, especially related to shellfish and marine food in general can have serious repercussions [11] and profit reductions due to a bad reputation with the general public. Likewise zones that are frequently detected with toxin levels above the acceptable levels will have their regional fish sectors prohibited from doing their work, hindering the mentioned sector as well as the resulting economy in that region, even if only temporarily [12]; in 2015, a record-breaking concentration of Domoic Acid was found on the western coastline of the United States, shutting down the fish and shellfish industries temporarily, leading to many (53%, n=197) in the fishing industry surveyed in Ekstrom et al. [13] declaring not being able to recoup their losses as of 2017 - roughly 2 years after the event. US Congress only supplied $26 Million USD to support the economic losses which was less than a quarter of the requested amount, which showcases the huge economic consequences HAB proliferation can cause [13].

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 [14]. 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 all 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.

1.2 Data Collection and Sources

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

NA - Não Aplicável NQ - Não Quantificado NR - Não Realizado

Nº Amostra	Espécie	Local de Amostragem	Zona de Produção	Data de Colheita	Toxinas Lipofílicas AO+DTXs (µg AO equiv/kg)	Toxinas Amnésicas (mg AD+AE/kg)	Toxinas Paralisantes (µg STX equiv/kg)
1715	Mexilhão	Jusante da Ponte	EMR	1/10/19	232	NQ	NQ
1721	Amêijoa-macha	Moacha	RIAV1	1/10/19	>550	NR	NR
1722	Berbigão	Moacha	RIAV1	1/10/19	267	NQ	NQ
1723	Amêijoa-japonesa	Canal do Espinheiro	RIAV3	1/10/19	233	NR	NR
1724	Berbigão	Canal do Espinheiro	RIAV3	1/10/19	455	NQ	NQ
1726	Amêijoa-branca	Torreira	L3	1/10/19	NQ	NR	NR

Figure 1.3: Small example of the data presented publicly by IPMA on their website about biotoxin rates in different shellfish in the various areas of the portuguese coastline. [2]

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 [16]. Copernicus will thus, be a valuable source of information to extract further data such as Chlorophyll and Sea Surface Temperature (SST).

Due to the high volume of data captured from these different sources, not all of them were used but due to the growing nature of this project and future work, they were still found worth mentioning.

2

Concepts

This thesis has a broad scope and is open to several different methods and applications that can be combined into the single overall purpose of forecasting hazardous biotoxin rates in shellfish in shoreline waters. As such, many concepts (from the basic time series and its properties to forecasting and classification methods) need to be properly introduced with, if possible, related papers and scientific studies about said concepts.

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 [17]. A common notation to represent TS is the following:

$$X = \{X_t : t \in T\}, \tag{2.1}$$

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 [18]. This differs from a Univariate Time Series where only one variable changes through time, as the name suggests [19].

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 [20]. 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) [21]. 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 [22]. 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 Lag

The n-th lag at a certain data point (at a specific time *n*) represents the data point observed at the moment r_n. The lag serves to allow the assessment of the evolution of certain attributes over time and study patterns in the time-series (such as the seasonality) as well as to enhance the forecasting accuracy of certain models with the help of other characteristics such as the trend. They are also crucial when studying certain analysis methods that depend on lagged data points like the Autocorrelation Functions (and its partial variant). [23]

2.5 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 [24]. Eq.2.2 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 $i \in T$.

$$r_k = \frac{\sum_{i=k+1}^{T}(y_i - y^{'})(y_{i-k} - y^{'})}{\sum_{i=1}^{T}(y_i - y^{'})^2} \tag{2.2}$$

2.6 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 [24]. Eq.2.3 showcases this mathematically.

$$r_k = \frac{\sum_{i=k+1}^{T}(y_i - y^{'})(y_{i-k} - y^{'})}{\sum_{i=k+1}^{T}(y_i - y^{'})^2 \sum_{i=k+1}^{T}(y_{t-k} - y^{'})^2} \tag{2.3}$$

2.7 Information Criteria Methods

2.7.1 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 [25]. 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:

$$AIC = -2\ln(L) + 2p, \tag{2.4}$$

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

2.7.2 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:

$$BIC = -2ln(L) + pln(n), \tag{2.5}$$

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 [26]. 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).

This model also aims for the model that minimises its criterion result. It has been attempted to combine these models but they have aspects that were impossible to reconcile, as Yang (2005) [25] showed. Studies to compare these were done and Acquah (2006) [27] and Markon et al. (2004) [28] have shown that AIC tends to perform better with small sample sizes but has inconsistencies and ends up performing similarly with bigger samples; BIC in contrast is more consistent and improves its performance with the increase of sample size. For this reason, AIC will be the preferred criteria used in model selection whenever possible.

2.8 Performance Measuring Methods

2.8.1 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:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2, \tag{2.6}$$

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 [29].

2.8.2 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:

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2]^{\frac{1}{2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2}. \tag{2.7}$$

2.8.3 Mean Absolute Percentage Error

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}(|\frac{y_i - y_i'}{y_i}|). \tag{2.8}$$

2.9 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:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + ... + \varphi_p x_{t-p} + a_t \tag{2.9}$$

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, named Intercept Term [30].

This can also be written as:

$$x_t = c + \sum_{i=1}^{p}(\varphi_i x_{t-i}) + a_t = \sum_{i=1}^{p}(\varphi_i x_{t-i}) + a_t \tag{2.10}$$

2.10 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:

$$x_t = c + \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_p a_{t-p} + \varepsilon_t \tag{2.11}$$

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 [31]. Like the AR model, it can be re-written as:

$$x_t = c + \sum_{j=1}^{q}(\theta_j a_{t-j}) + a_t \tag{2.12}$$

2.11 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:

$$x_t = \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{2.13}$$

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 [32].

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:

$$x_t = \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) + \sum_{j=1}^{0}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{2.14}$$

$$= \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) + \varepsilon_t \tag{2.15}$$

$$= \delta + \sum_{i=1}^{p}(\phi_i x_{t-i}) = AR(p) \tag{2.16}$$

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

$$x_t = \delta + \sum_{i=1}^{0}(\phi_i x_{t-i}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{2.17}$$

$$= \delta + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \tag{2.18}$$

$$= \delta + \sum_{i=j}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t = MA(q) \tag{2.19}$$

2.12 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 [32]. An ARIMA model is typically described as ARIMA(p,d,q) as it showcases all the elements of the model:

- AR (Autorregression) - the regressive model shows the variable changing by regressing on its own lagged observations (p).
- I (Integrated) - applies differencing between the values in order to allow the series to become stationary (d).
- MA (Moving Average) - takes into consideration the dependency between observations and a residual error from a MA model applied to lagged observations (q).

An ARIMA model ARIMA(p,d,q) is then written as such:

$$x_t - \alpha_1 x_{t-1} - ... - \alpha_p x_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-1} \tag{2.20}$$

Which is equivalent to:

$$(1 - \sum_{i=1}^{p} \alpha_i L^i) x_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \epsilon_t \tag{2.21}$$

2.13 Random Forests

Random Forests (RF) are an ensemble learning algorithm that build a set of decision trees that are then trained and are then used for classification or regression. They originated in 1995 by Tin Kam Ho and [33] were then extended and popularized through Breiman [34]. Breiman's algorithm complemented Ho's through the introduction of bagging and random feature selection - by training each tree with a random set of data samples, the learnt results are the multiple uncorrelated trees built during training [34]. 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 [35]. The learnt smaller models (the trees) are then combined into a single prediction result using Breiman's bagging idea - by combining the predictions given by each tree, the final result will compile them and apply a method to the returned combination of results to give the overall prediction value [34]. 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. Naturally this leads to an extrapolation problem - because the values in all leaves are averages of previously seen samples/observations, Random Forest Regressors cannot predict and extrapolate values outside of the range present in the training set as the resulting values (for both the trees and the Forest itself) result from an average, which can never sit outside of the set's minimum and maximum values; this consequence leads to its inability to predict possible trends that put the value range outside of the training set's range so choosing to use RF's can have dire consequences if the data samples prove to be too volatile in its value ranges over time as there's no guarantee those values stay the same in the future [36]; in contrast, its ability to minimize overfitting was a strong argument for its usage in this work. The motivation to use RF's in this work was also helped by existing research in this theme - Cheng et al. [37] used an Interative Random Forest (iRF) [38] 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. This iRF algorithm aims to grow iteratively an N number of re-weighted RF's. This re-weighting is done through the Gini Importance (GI) index, which measures and re-calibrates feature-importance during the decision process. After obtaining the decision rulings from the built RF outputs, the generalization of the built trees is done and results in the above mentioned re-weighted Random Forests. The iRF used in this study was applied to a Santa Cruz Wharf weekly dataset that ranged from 2011 to 2019 with nearly the entirety of 2018 dismissed due to missing values. 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). There were also detected quantifiable stable interactions between Phytoplankton OTU's (Operational Taxonomic Unit), among them the *Pseudo-nitzschia* group. Other RF studies also proved fruitful: Valbi et al. [39] 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 (among them nutrients, SST and salinity), 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 the nutrient features of the dataset were discarded, reducing the dataset attributed from 18 to 12. This new study had more variability in the predictions but the best model with the smaller subset of variables was just as good and even slightly better than the best model obtained with the full feature set - this 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 (since nutrient samples weren't needed anymore) and its variables were computed using the z-score (calculated by dividing their raw scores by the respective standard error [34]) to determine their importance for the studied model.

2.14 Bayesian Networks (BN)

BN's are statistical models that represent attributes and their conditional dependencies in a directed acyclical graph (DAG) [40]. 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 [41]

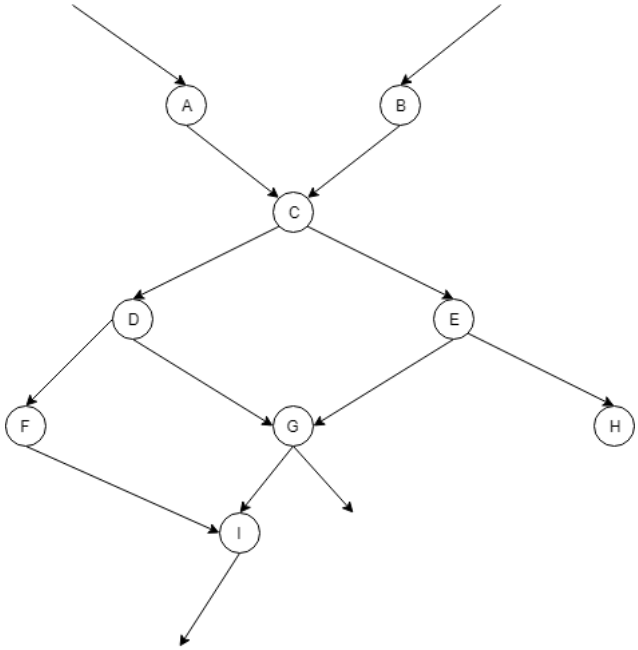

Figure 2.1: An example of a portion of a Bayesian Network Graph.

From Figure 2.1, we can conclude that D and E are conditionally independent, given C. They are not totally independent as they have originated from the same variable (C), but the expressions to obtain both D and E and different from one another. For this graph, we have a set of variables, represented by nodes $N = \{A, B, ..., I, ...\}$ which are connected by edges (representing the dependencies) in a set *E* and a set *P* representing the probability distribution function for each variable in *N*.

Taking the example with nodes C, D and E again, 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 must be known to calculate any inferences related to these attributes.

To give a simple example using the Figure 2.1 above, we can write the Bayes rule of posterior probability, $P(D|C)$, given P(D) and the likelihood $P(C|D)$:

$$P(D|C) = \frac{P(C|D)P(D)}{P(C)} \tag{2.22}$$

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

$$P(D|C) = \frac{P(D,C)}{P(C)} \tag{2.23}$$

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

$$P(D|C,E) = P(D|C). \tag{2.24}$$

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).

2.15 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 [42] [43]. 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

$$P(x_t|x_{t-1}) \tag{2.25}$$

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

$$P(X_{1:T}) = P(X_T|X_{1:T-1}) \tag{2.26}$$

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:

$$P(X_{1:T}) = \prod_t \prod_i P(X_{t,i}|pa(X_{t,i})) \tag{2.27}$$

Where i groups variables in a same time slice and pa(X) represents the parents of X in the network [40]. 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 will be used for the context of the data analysis required in this thesis.

2.16 Artificial Neural Networks

Artificial Neural Networks (ANN) are an architecture loosely inspired by how the brain works through neurons and their connectivity and characteristics [44]. 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. The output of a neuron travels through an edge and becomes the input of the receiving neuron, until the final output is emitted. 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 - Figure 2.2 represents an ANN described in Jain et al. (2006) [3] and perfectly showcases the above mentioned three layers.

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.

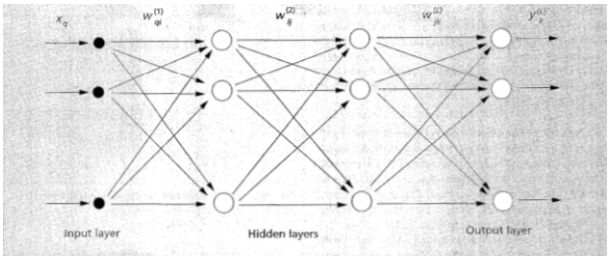

Figure 2.2: Typical representation of an ANN [3]

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 [45]. Three lakes and one river were studied, each with their own ANN system based on the time series data for each; 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 [45].

2.17 Gradient Boosting and XGBoost

2.17.1 Gradient Boosting

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 [46]. 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) [47] studied Friedman's Gradient Boosting documentation [46] 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:

$$y_i' = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{2.28}$$

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

$$L = \sum_i l(y_i', y_i) + \sum_k \Omega(f_k), \tag{2.29}$$

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:

$$\Omega(f) = \gamma T + \frac{\lambda w^2}{2}, \tag{2.30}$$

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. [48]) and classification, such as Torlay et al. [49] which managed to classify patients with epilepsy with an AUC (Area-Under-Curve) mean score of 91%.

3

Time Series Analysis

3.1 Challenges

The process of data collection proved to be a bigger challenge than expected and took a considerably larger amount of time to get it to a valid state ready to train and evaluate and (consequently) forecast.

The data provided is publicly available at ”https://www.ipma.pt/pt/bivalves/docs/index.jsp” and each data file has all samples collected in a given month of a given year; these samples are ordered by date and contain all species and respective zones of collection. Species differ depending on the data collected: biotoxin files are made of shellfish species and their respective contamination values of DSP (Diarrhetic), ASP (Amnesic) and PSP (Paralytic). Phytoplankton data files are comprised of cell counts (through the Utermohl method) in a water sample.

The data files were originally in a PDF format and thus, a conversion to a Comma Separated Value - CSV format was chosen due to its readability and versatility when using different libraries and languages.

Furthermore, 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):

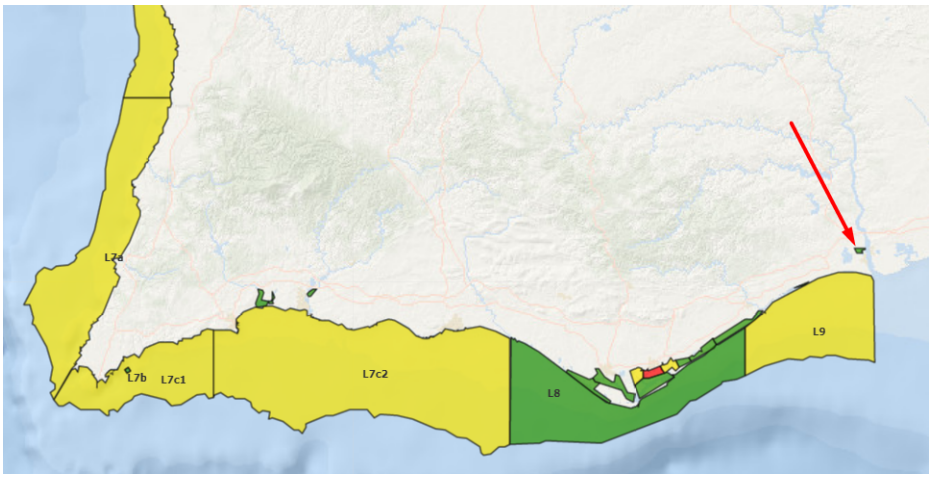

Figure 3.1: GUA region (small region pointed by the red arrow), one of the new zones added halfway into the analysis process throughout the years.

- Zones changed names over the years and some were even introduced throughout the years. The GUA zone is an instance of this as it was only introduced in 2017 and is situated very closely to the Guadiana river in the south of Portugal - see Figure 3.1. This is logically also present in the phytoplankton data and had to be accounted for too.

- Toxin value thresholds changed over time - earlier in 2015/2016 being 850 (any value above it was simply referenced as $850\ \mu g$ 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.

  These changes proved an interesting challenge to tackle as the series used in this work needed more pre-processing to be homogenized and better prepared to be studied by the models developed, though there is room to further explore these challenges in other ways in future work.

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.

3.2 Joining the Data

To get the data to a stable and consistent point ready to be analysed, trained and forecasted, several procedures had to be done first to allow it to be joined into a single DataFrame ready to be pre-processed and worked with afterwards.

Firstly, the ND and NQ values were replaced by different values, depending on the toxin measured: for ASP toxins, the value *1.8* was considered, *28* for DSP toxins and *71* for PSP toxins. These were the recommended values to be used for the time series when IPMA were inquired about the values ND and NQ took in their analysis.

NR values adopted a missing-value approach and as such, different methodologies could be used. For this work, three simple approaches were used - the first two were using the mean and median of the remaining values of the time series dataset, while the third and final one used the mean of the values of the two closest data points.

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.

For biotoxin data, the zones in older data files were changed to their new names and the new ones were also included (a time series in that zone will obviously have less data points on average). As for contamination thresholds, values above them were not explicitly input so they were replaced by their threshold value (for instance, $>$*850* was replaced by *850* and this for each respective threshold change (meaning the same procedure was done for thresholds *625* and *550*).

For phyotplankton data, the changes in analysed species proved easy to homogenize as only DSP, ASP and PSP values were to be accounted for in this work so all other columns were removed and as a consequence, were not a part of this study. Furthermore, the procedure was similar to the one done for the categorical data in the biotoxin dataset: after inquiring about this change in the data starting mid-2018, the returning information recommended that the previous zero values should be instead considered *Below Threshold* as well and afterwards converted to a value of 20.

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.

The Copernicus data is presented in a NetCDF format which is rich in information but requires packages to be read and processed - for this, the Python **xarray** package was used. The time series could then be extracted and even visualized as the Copernicus data always contains the geographical coordinates of the collected data, thus allowing their SST and chlorophyll data to be adequately added to the zone-specific time series in this work (see Annex B).

	Colheita	Amostra	Espécie	Local de Amostragem	Produção	TL	TA	TP	CHL	SST (celsius)
0	2016-01-18	43	Conquilha	Culatra	L8	118	ND	ND	0.49	15.869000
1	2016-01-26	117	Conquilha	Culatra	L8	84	ND	NQ	0.44	15.706000
2	2016-02-02	150	Conquilha	Culatra	L8	127	ND	ND	0.43	15.984000
3	2016-02-03	311	Conquilha	Culatra	L8	NQ	ND	NQ	0.42	15.899000
4	2016-02-06	960	Conquilha	Culatra	L8	241	NR	NR	0.77	16.322001

Figure 3.2: First 4 data points of the L8 Wedge Clam dataset, including the SST (^{o}C) and Chlorophyll values (mg/m^3), the final 5 attributes are the ones being studied and are, from left to right, the three studied toxins (Lipophilic, Amnesic and Paralytic), Chlorophyll and the SST.

After this brief description of the changed made to have a consistent dataset (the first values can be seen in Figure 3.2), pre-processing and analysis could begin.

3.3 Pre-Processing

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. The missing value rate is presented in Table 3.1 for all 3 toxins that IPMA is currently providing information for.

	Toxins		
	Lipophilic	TA	TP
Missing Values (%)	2.9	17.6	21.8
Data Points (total)	9136		

Table 3.1: Overall statistics of the Time Series data with the percentage of missing values for each toxin.

This image only serves to represent overall data because there is no logic is working with a dataset containing all species and all areas at the same time. As such, the pre-processing and overall analysis process will be done to smaller datasets that encapsulate a single species on a specific capture area.

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 as Figure 3.5 showcases, proving to be much less fruitful datasets to work with.

The Copernicus data was also plotted to see how values evolved over time. Figures 3.3 and 3.4 showcase this.

	Copernicus Data	
	SST (ºC)	Chlorophyll (mg.m^{-3})
Count (Data Points)	166	166
Mean	17.58	0.95
Standard Deviation	2.46	0.76
Minimum Value Recorded	13.94	0.05
Maximum Value Recorded	23.97	4.26

Table 3.2: SST and Chlorophyll general statistics over the 4 year span study of the L8 area.

Using the time series statistics present in Table 3.2 as a starting point for further studies, the Wedge Clam information on the L8 area will be used for the forecasting study of this dissertation, among a few other series used for further analysis. As such, it's worth mentioning the overall statistics of this dataset, similarly to Table 3.3:

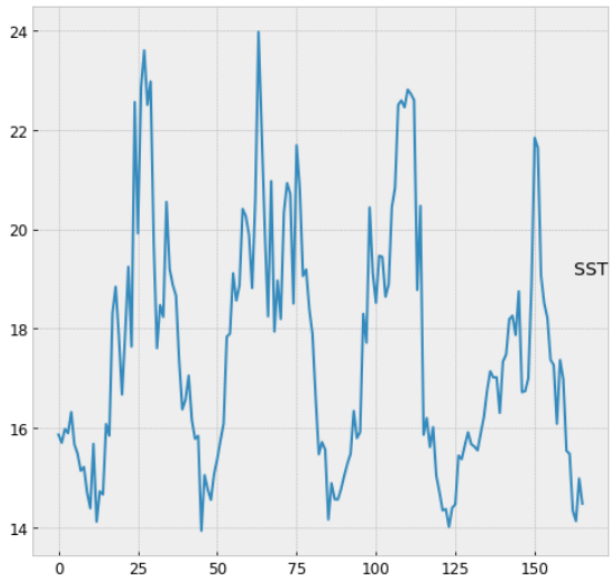

Figure 3.3: Sea Surface Temperature (degrees celsius) in the L8 area across all datapoints.

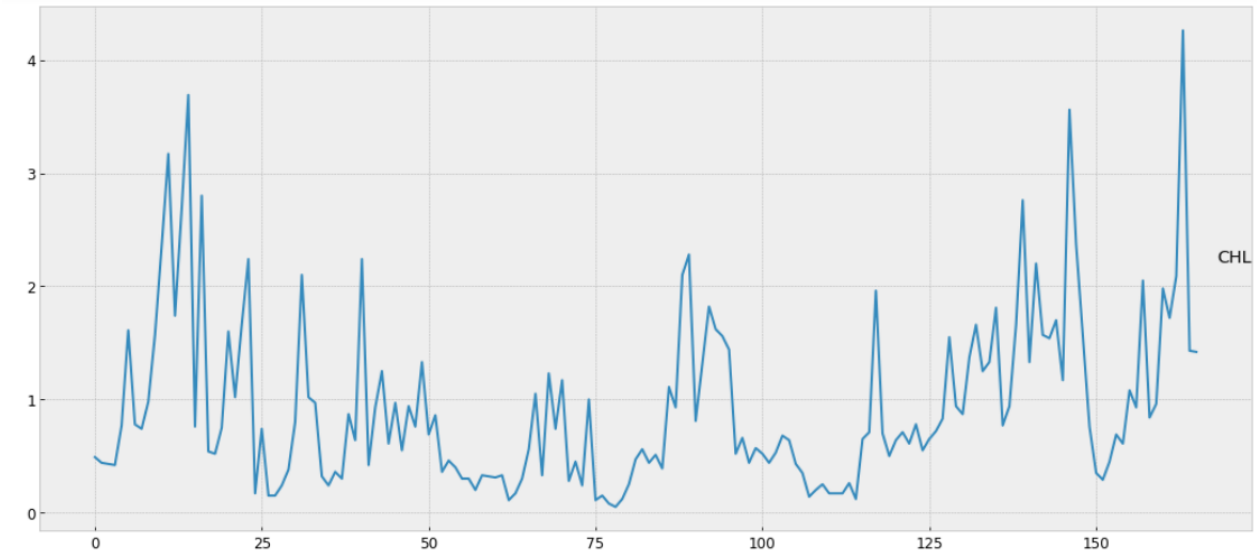

Figure 3.4: Chrolophyll (mg/m^3 - milligrams per cubic meter) values detected in the L8 area across all datapoints

	Toxins		
	Lipophilic	TA	TP
Count (Data Points)	166		
Mean	204.50	2.16	83.22
Standard Deviation	181.70	1.73	111.68
Minimum Value Recorded	21	1.8	11
Maximum Value Recorded	850	16	1491

Table 3.3: Overall statistics of the L8 Wedge Clam dataset. There are a grand total of 166 data points across the 4 years studied.

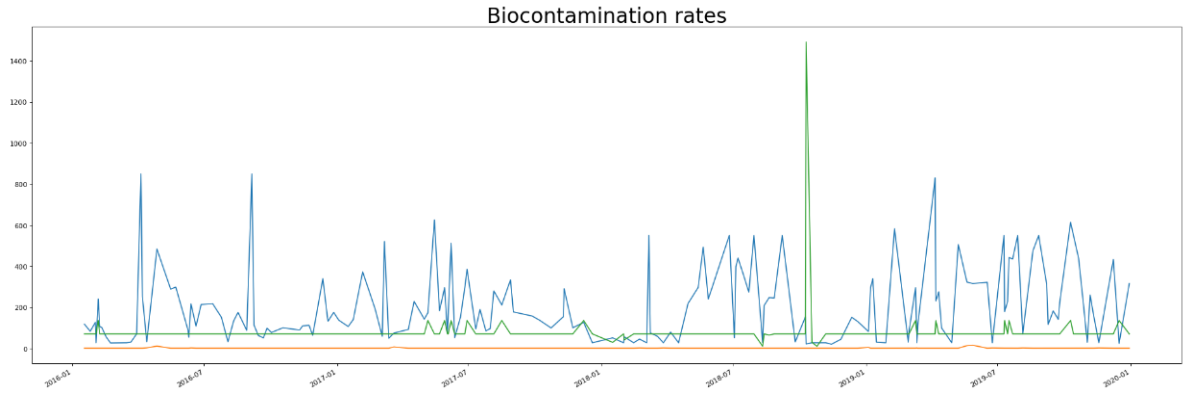

Figure 3.5: Toxin rates in Wedge Clam shellfish in the L8 area. Lipophilic can be see in blue, Amnesic in green and Paralytic in orange.

From this point onward, Python packages proved useful thanks to their added functionality that enabled studying other, more specific components of the dataset.

Using the **statsmodels** package, a function titled *seasonal decompose* was made available to decompose any time series into its various components. When applying it to the above mentioned dataset, the time series values were decomposed as shown in Figure 3.6.

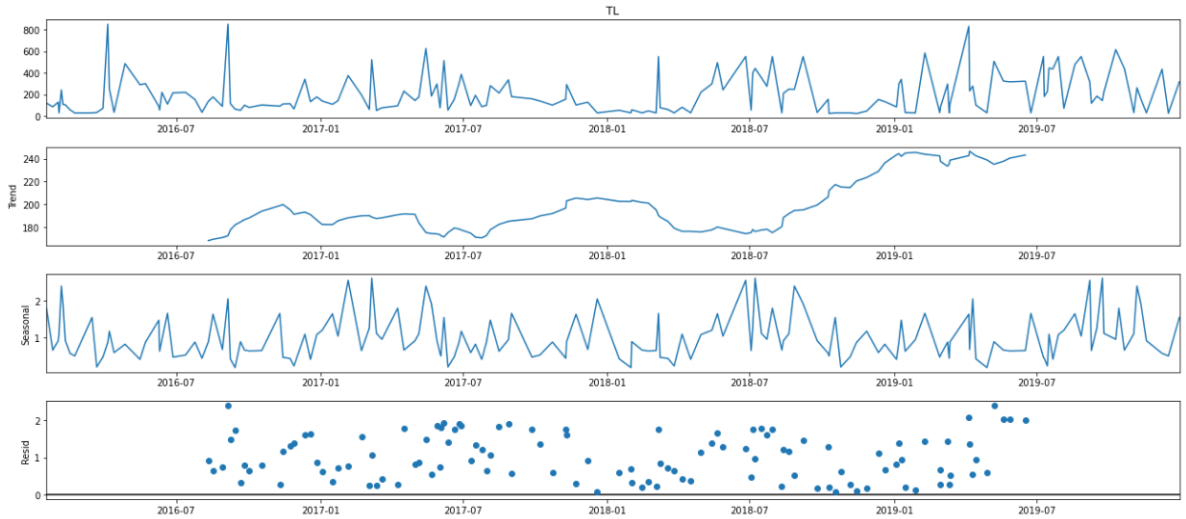

Figure 3.6: Decomposition of the Wedge Clam collection in the L8 area time series. The Series is decomposed into the Original value (at the top) and respectively lowers into the Trend, Seasonality and Remainder components.

We can see that the Trend was relatively consistent throughout the first three years, noticeably rising at the end of 2018 and staying high throughout 2019 which fits the noticeably higher average values witnessed throughout the same time periods in the original data.

There is no seasonal pattern in the series, however, as there does not seem to be a consistent pattern among the value variations in the decomposed section that represents the seasonal component

of the time series.

After this analysis, it was important to check on how attribute values correlated with lagged versions of themselves and thus better comprehend models to be applied and study the series. For this purpose, Autocorrelation and Partial Autocorrelation plots were generated to better see this.

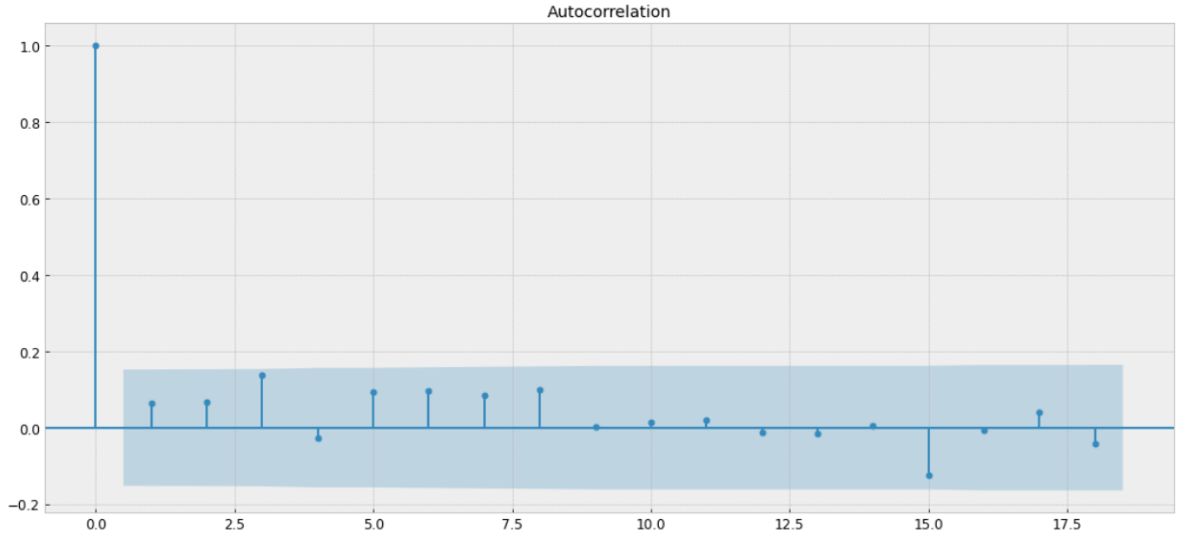

Figure 3.7: Caption

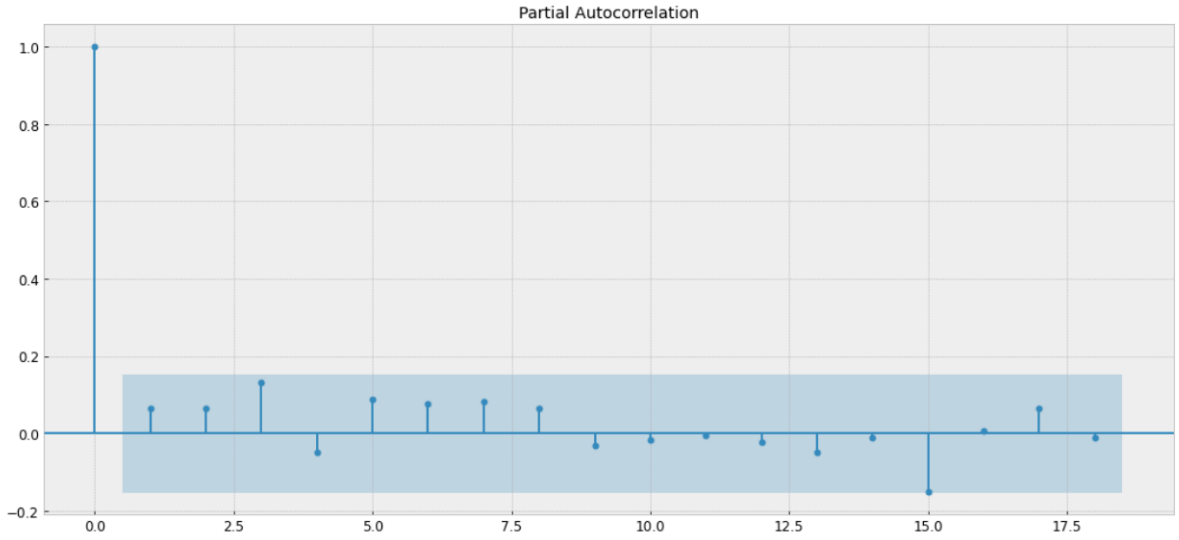

Figure 3.8: Caption

As we can see, there is not a very significant correlation between values and their lagged versions. At point zero, the value is 1 as the correlation is measured between the value and itself, meaning it is obviously the maximum achievable. The same applies to the Partial Autocorrelation plot. The first 7 lagged values seem to be consistently the highest but it is still a very small value - there is no real

exponential or consistent drop of any sort to represent a significant autocorrelation.

Another concept important in time series analysis that was discussed in the previous chapter is stationarity. For this, an Augmented Dickey-Fuller (ADF) test was made through Python's *adfuller* function of the **statsmodels** package. This Unit Root test was preferred to the other typical test, Phillips-Perron, due to the latter's tendency to underperform in comparison with ADF in finite time series samples (like the ones being dealt with in this work).

After this analysis and pre-processing phase, we will discuss model results in the next chapter for this dataset.

4

Results and Evaluation

4.1 Wedge Clam L8 Dataset - Forecasts

The Wedge-Clam L8 dataset 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.

In this chapter, a brief explanation of how the models were used (language, libraries/packages and functions, plus their respective main parameter values) and an image to showcase the final result on the forecasting of the above-mentioned dataset. In the end of the chapter, a table and plot of the evaluation metrics will be presented (previously described in Chapter 2). Overall conclusion and result discussion will be in Chapter 5.

4.1.1 Autorregressive - Auto-ARIMA Model

For the ARIMA model development, due to the several values to be tuned, a package named pmdarima was used - it wraps the classic **statsmodels** library, enabling easy usage for Python developers. Specific

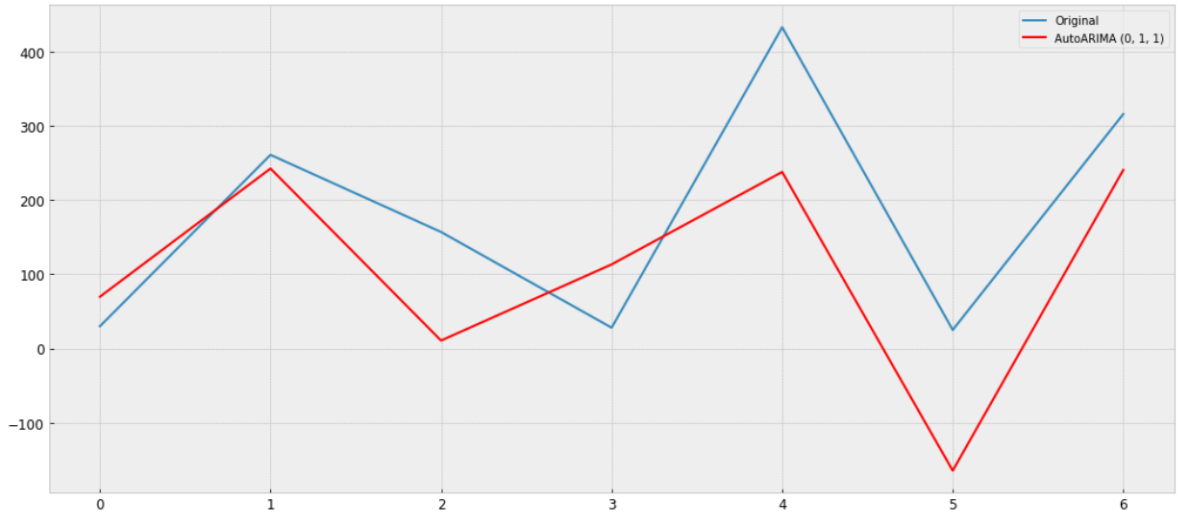

Figure 4.1: ARIMA performance on the Wedge Clam-L8 time series.

to this package is the *auto_arima* function that automatically picks the best version of the ARIMA model which was (0,1,1) for the respective (p,d,q) parameter tuple. This function was run with both the AIC and BIC information criteria methods - both yielding the same ARIMA model.

4.1.2 Random Forests Regressor- RF

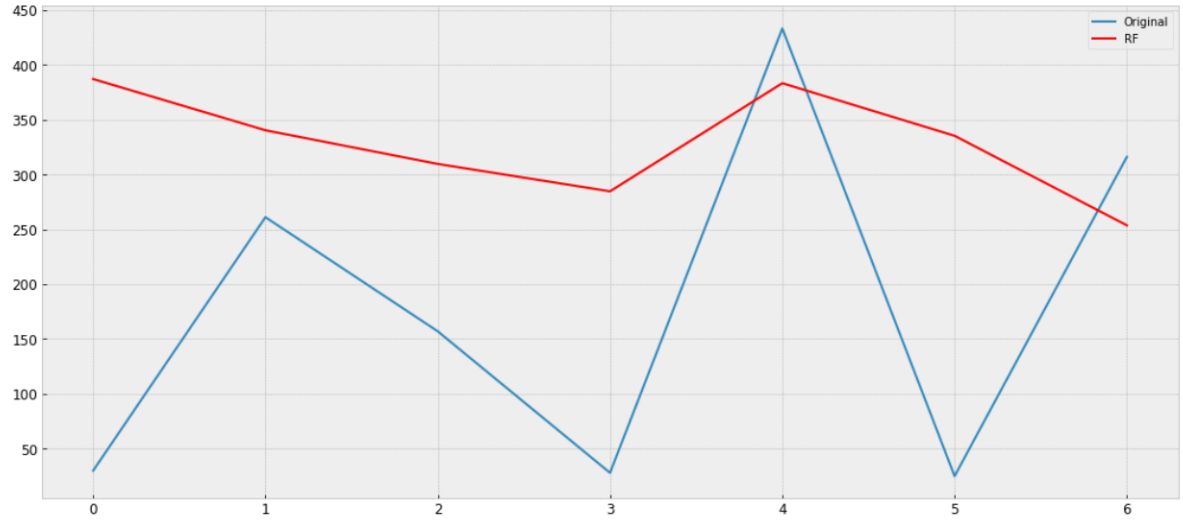

Figure 4.2: Random Forest Regressor applied to the Wedge Clam-L8 time series.

A Random Forest Regressor was implemented through the **sklearn** Python package. The most notable parameters that were tuned were:

- *n_estimators*: the number of trees built by the regressor - this was picked as 100 due to it being a

small value, fast to compute and because biggers values were tested and proved to yield little to no improvement - this is consistent with the fact that the number of trees helps improve the model until a point where the error rate improvements are negligible and are not worth the exponential performance increase in generating and ensembling all the trees.

- *max_depth*: chosen as 2 to allow the model to be flexible with data. Large values of tree depth make it prone to overfitting (despite Random Forest's randomness in tree generation and averaging) and there was also no noticeable improvement after a substantial increase in this value.
- *n_jobs*: no influence in the model itself - it unlocked all processor cores for usage to achieve faster processing and model training.

4.1.3 Gradient Boosting Trees - XGBoost

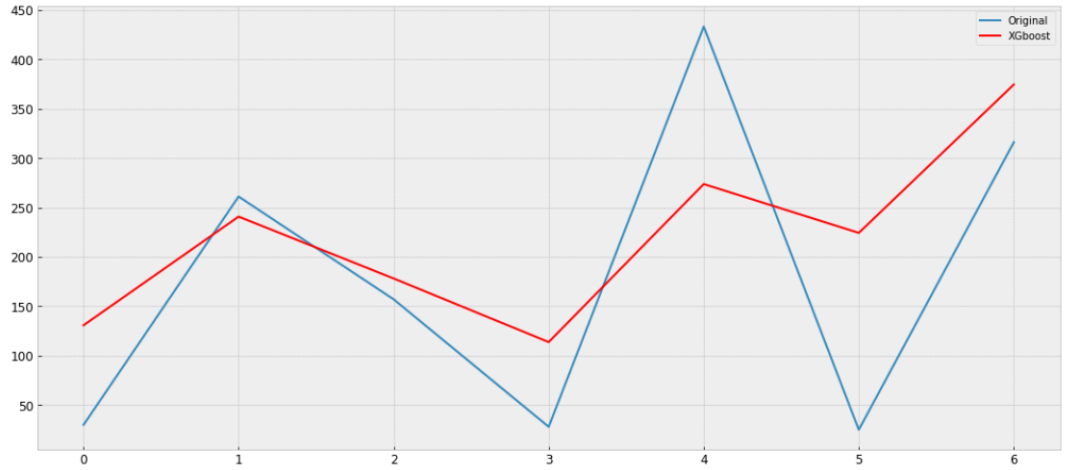

Figure 4.3: XGBoost forecasting performance on the dataset.

For the XGBoost forecasting, a similar approach was made to the Random Forest Regressor and the following parameters were selected (some of them were the default values but those ended up being logically reasonable ones):

- *n_estimators*: taking the value of 100 for the precise same reasons as the Random Forest counterpart - even with the new features associated with XGBoost, there were no noticeable improvements past 100.
- *objective*: a string value was picked *squarederror* meaning that the learning objective to be minimized was the MSE metric already approached in Chapter 2.
- *gamma*: the larger this parameter is, the more conservative the algorithm would be and thus, within the allowed range of $[0, \infty]$, the minimum value of 0 was used.

- *max_depth*: after testing various combinations (and also using a Python package names **BayesianOptimization** to facilitate this testing), the best result was yielded with a *max_depth* value of 4 without showing overfitting.

- *booster*: a string value was used (*gbtree*) to make the model use a tree based booster as was explained in Chapter 2 (other values such as *gblinear* are possible too).

4.1.4 Evaluation Through Metrics

For this select dataset, the follow evaluation metrics for the models were obtained:

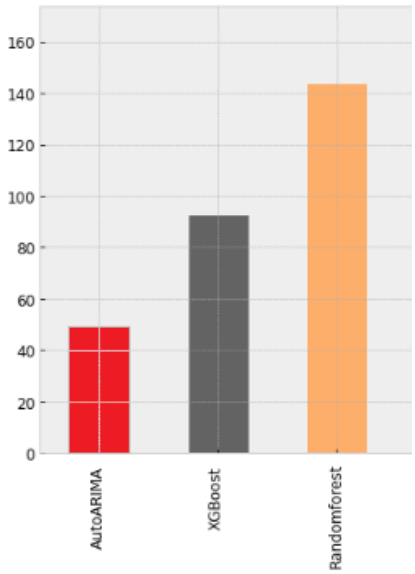

Figure 4.4: MAE metric results for each model trained with the Wedge Clam dataset on the L8 area.

Figure 4.4 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.

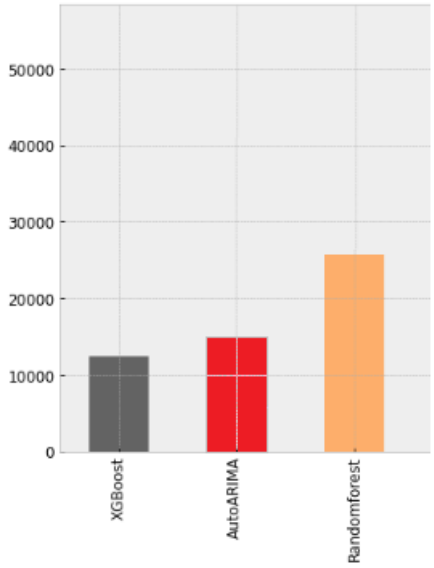

Figure 4.5: MSE metric results for each model trained with the Wedge Clam dataset on the L8 area.

As Figure 4.5 showcases, the AutoARIMA and XGBoost models obtained a respectable value that complements their good performance on the MAE evaluation.

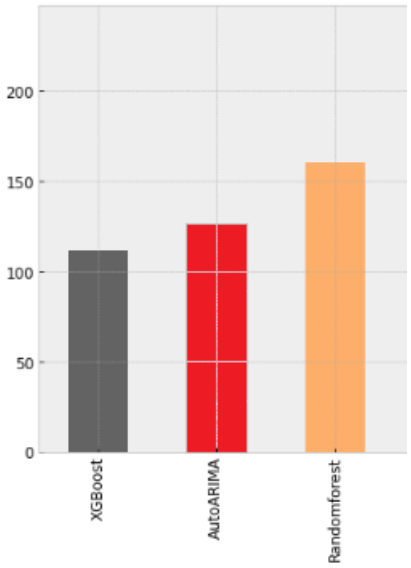

Figure 4.6: RMSE metric results for each model trained with the Wedge Clam dataset on the L8 area.

Finally - and similarly to the MSE results - the RMSE performance was better with the XGBoost and AutoARIMA models.

4.2 Complementary Study (including Multivariate Time Series) using MAESTRO

For the following analysis, 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 - with its respective location seen in Figure 4.7.

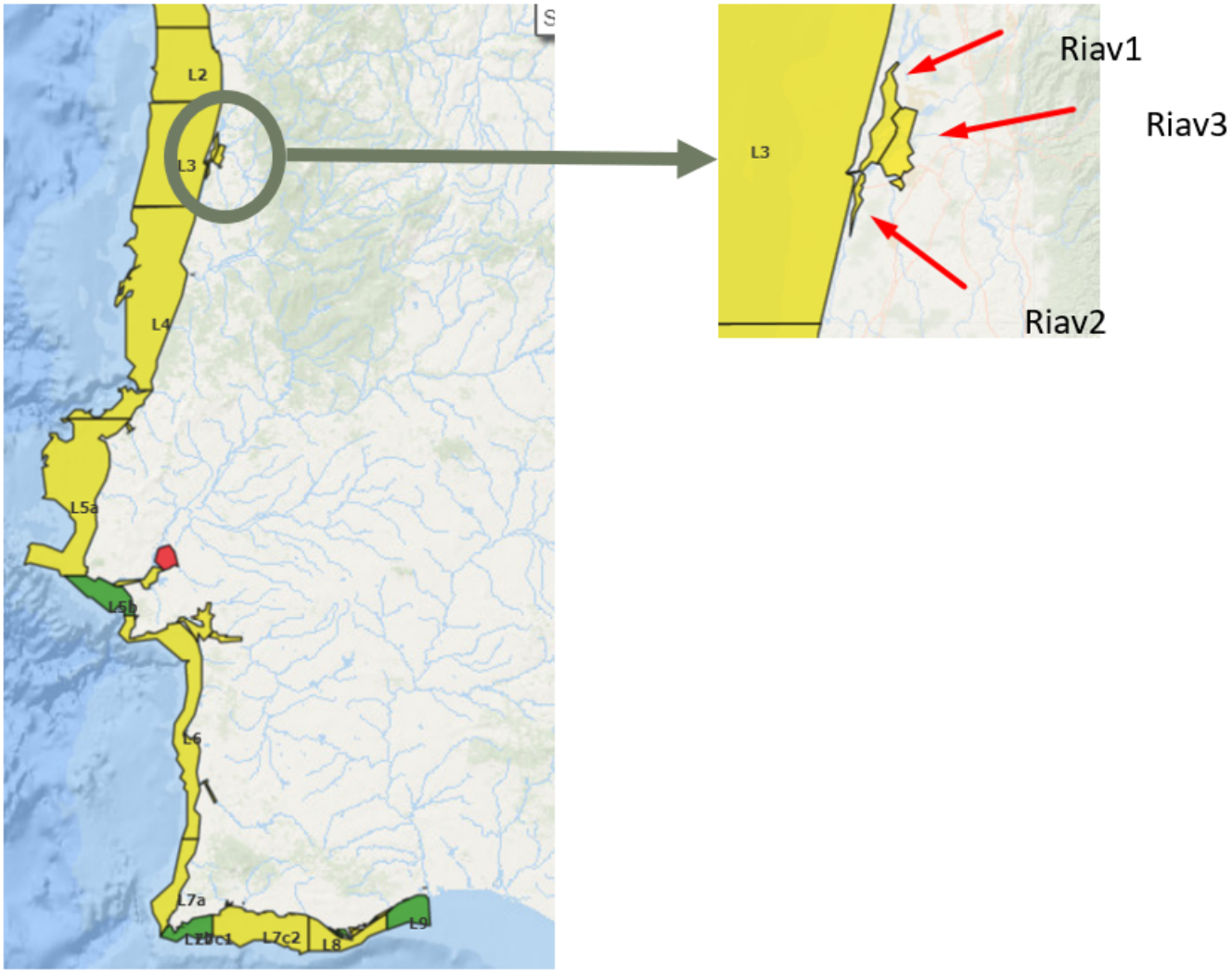

Figure 4.7: Location and relative position of RIAV subgroups in Portugal's Coastline

The reason RIAV 4 was not taken into account can be justified by looking at Appendix D, where the number of time series datapoints are very similar among the first three RIAV zones, but the fourth possesses far fewer samples, making its inclusion in this analysis very complicated and nigh impossible

due to the sample size disparity that would likely cause more error-prone results.

For the RIAV datasets, 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).

Because this subsection approaches a multivariate approach, the goal in this section was to progressively add more variables in order to see the evolution of MAESTRO's generated networks and make a final assessment of how variables interact with each other within a single time-series and also between different time series. For this, a series of plots will be shown of each attribute of the time series, alongside the MAESTRO counterpart, concluding with the MAESTRO-generated Dynamic Bayesian Network modeled (and respective conditional probability tables in the Appendix sections). Once all RIAV zones are explored, a study of how a time series comprised of two of these zones combined behaves under MAESTRO's Dynamic Bayesian Network modelling will be discussed.

4.2.1 RIAV1 Dataset Presentation

The images below show MAESTRO's attribute visualization for the RIAV1 dataset with 5 variables where red represents the lowest values and blue the highest, the colors change depending on each variable's value range. To ease the explanation of the following images, they are complemented with the respective variable plot. By carefully looking at both plots of the same, variable, it can be seen that, for instance, blue datapoints in MAESTRO coincide with the peaks seen in the respective plotted variable.

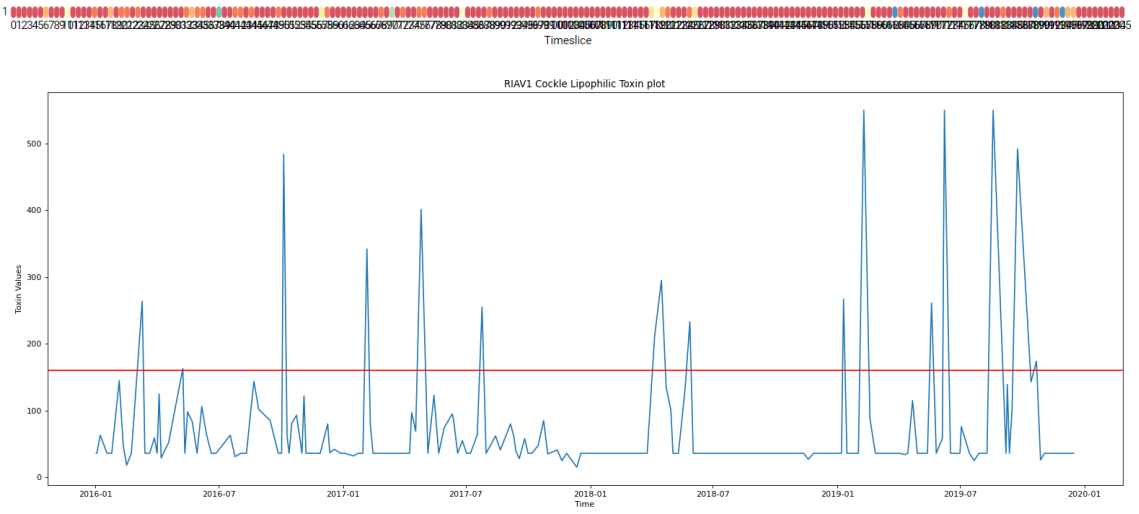

Figure 4.8: Vertical comparison of the lipophilic biotoxin data of the RIAV1 dataset, MAESTRO (above) and a built plot with the interdition restriction threshold in red (below).

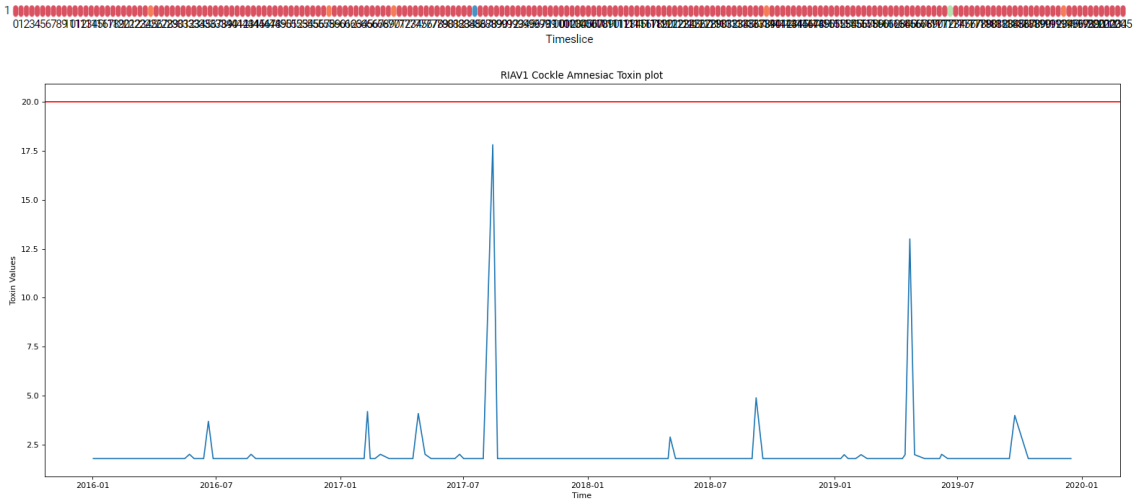

Figure 4.9: Vertical comparison of the amnesic biotoxin data of the RIAV1 dataset, MAESTRO (above) and a built plot with the interdition restriction threshold in red (below).

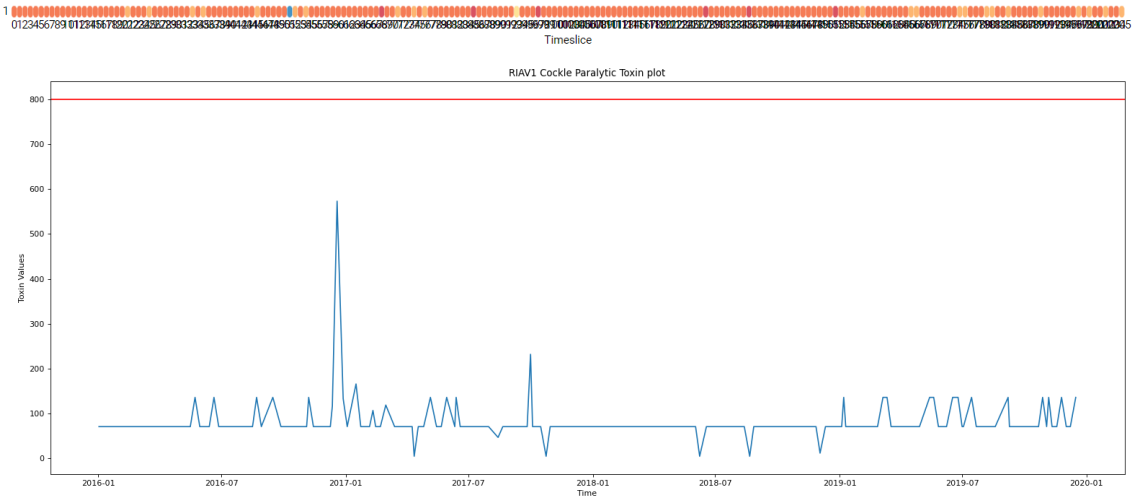

Figure 4.10: Vertical comparison of the paralytic biotoxin data of the RIAV1 dataset, MAESTRO (above) and a built plot with the interdition restriction threshold in red (below).

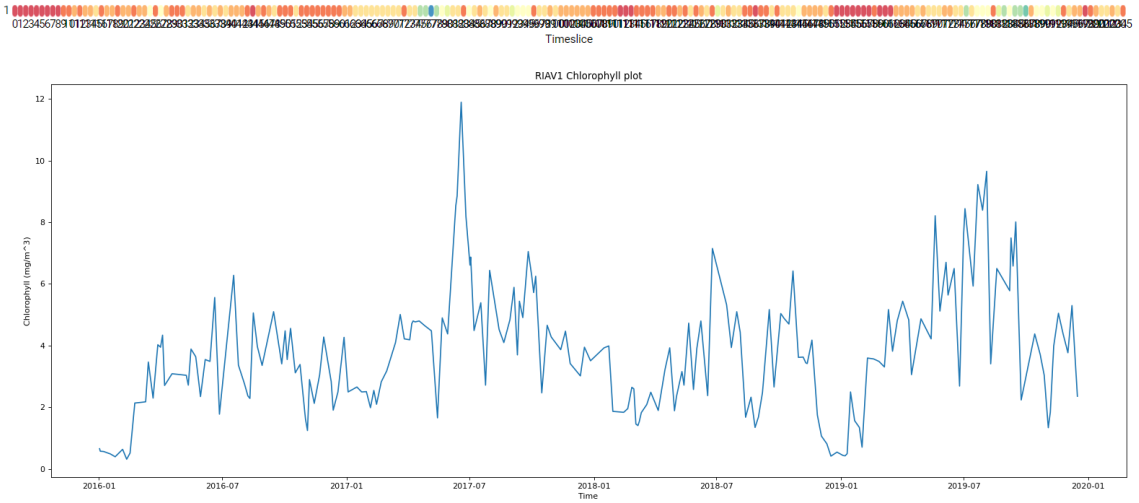

Figure 4.11: Vertical comparison of the chlorophyll data of the RIAV1 dataset, MAESTRO (above) and a built plot (below).

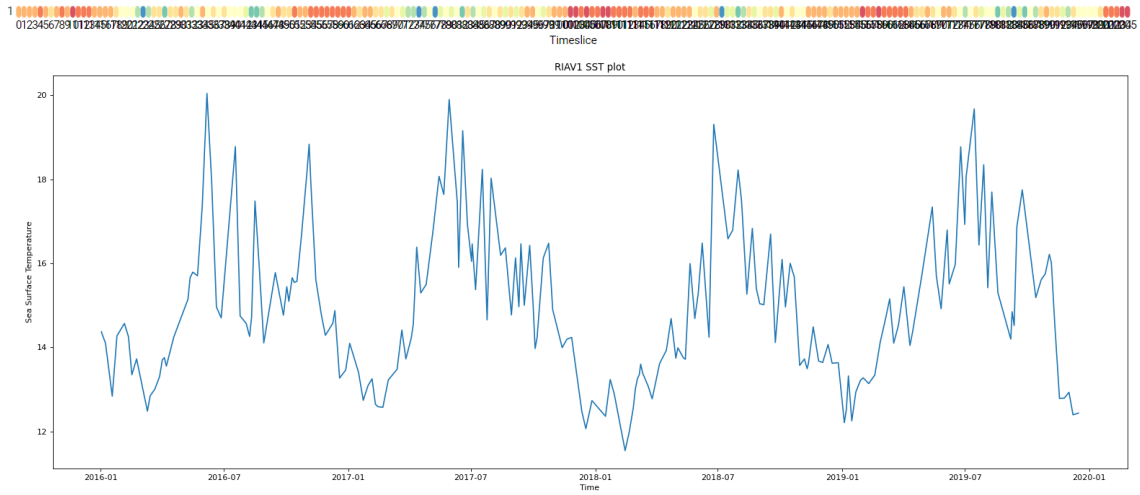

Figure 4.12: Vertical comparison of the sea surface temperature data of the RIAV1 dataset, MAESTRO (above) and a built plot (below).

As observed in Figure 4.13 , due to the high rate of values that were missing (NR) or below-threshold (ND and NQ), the data values suffer little variation throughout the 4 year span of the dataset, as the plots show (and MAESTRO's color plot, where most red datapoints represent the threshold values that were imputed to replace ND/NQ values). Whenever there is a change in the recorded values, MAESTRO's data records a different color (though small value fluctuations yield similar colors). This gives a better perspective of how MAESTRO uses the time series to create its models as the time series values have to be discretized - something this application does automatically for the user. The same figure also exhibits the fact that PSP and ASP toxins in particular can be seen having values under the detection or

quantifiable threshold combine for a huge portion of the time series (in the case of ASP, 91.3% for RIAV1, 87.1% for RIAV2 and 90.4% for RIAV3), something that can be corroborated by consulting Figures 4.9 and 4.10where the plot showcasing biotoxin concentrations in the Cockle shellfish very rarely fluctuate above the minimum threshold of detection/quantification.

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

	RIAV1 (569 datapoints)		
	DST Phytoplankton	AST Phytoplankton	PSP Phytoplankton
<LD	336 (59,1%)	372 (65,4%)	506 (88,9%)
	RIAV2 (312 datapoints)		
	DST Phytoplankton	AST Phytoplankton	PSP Phytoplankton
<LD	161 (51,6%)	189 (60,6%)	272 (87,1%)
	RIAV3 (323 datapoints)		
	DST Phytoplankton	AST Phytoplankton	PSP Phytoplankton
<LD	221 (68,4%)	217 (67,2%)	284 (87,9%)

Figure 4.14: Below Threshold (LD) value counts in the RIAV1, RIAV2 and RIAV3 time series, with respective rate percentage.

Afterwards, RIAV1 was complemented with the phytoplankton data, adding 3 new variables to the dataset for a total of 8. Because phytoplankton and toxin data are almost always collected in different days of the week (sometimes in different weeks), the opted procedure was to join both dataset's values on the nearest respective date for each row - with a one week threshold (more than a week did not seem productive for the bayesian network inference). Because some date differences exceeded this threshold, imputation of missing values was needed - MAESTRO also provided assistance and the missing values were filled using its *Last Observation Carried Forward* method.

The resulting data is presented in Figures 4.15 - 4.17 for the 3 phytoplankton data collections (DSP, ASP and PSP):

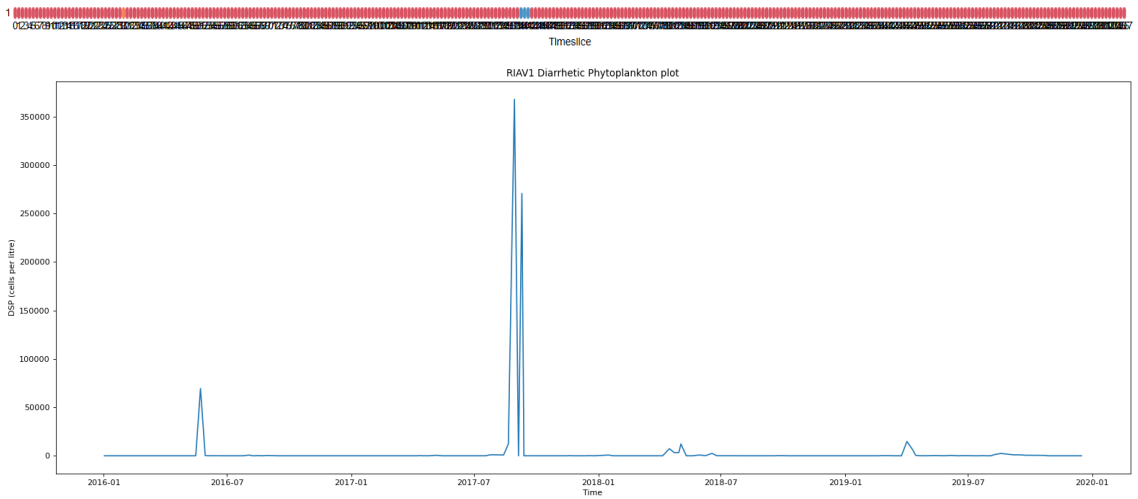

Figure 4.15: Vertical comparison of the DSP producing phytoplankton data of the RIAV1 dataset, MAESTRO (above) and a built plot (below).

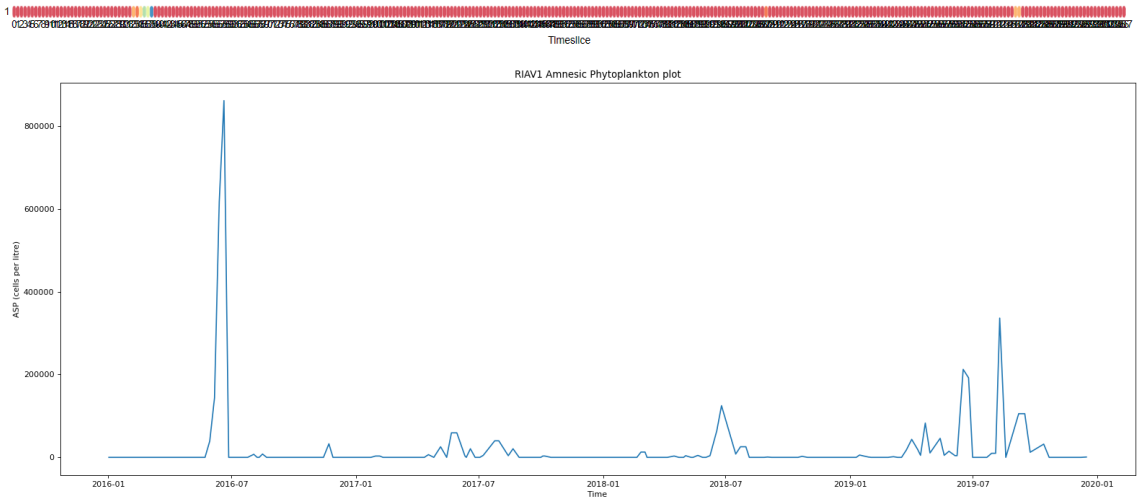

Figure 4.16: Vertical comparison of the ASP producing phytoplankton data of the RIAV1 dataset, MAESTRO (above) and a built plot (below).

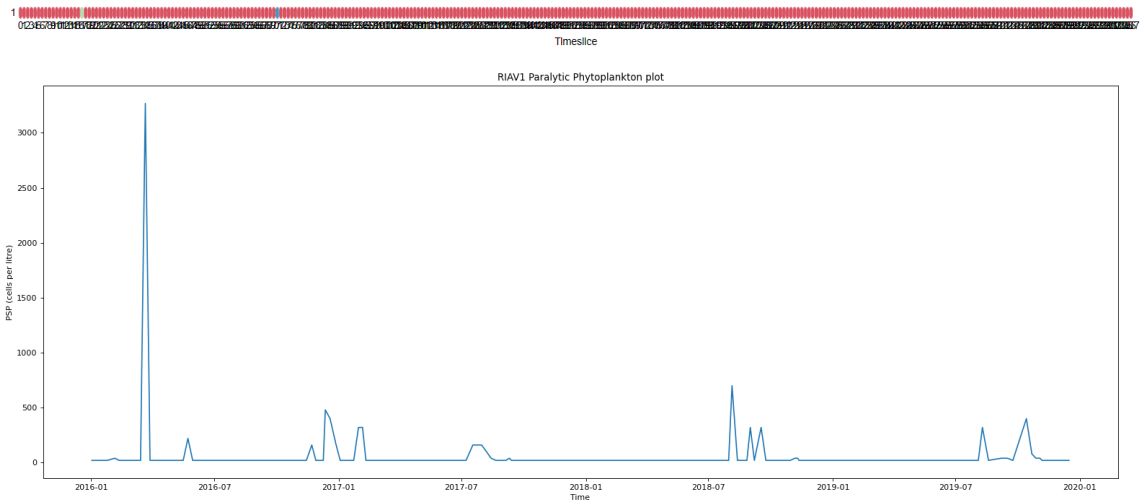

Figure 4.17: Vertical comparison of the PSP producing phytoplankton data of the RIAV1 dataset, MAESTRO (above) and a built plot (below).

With the added data, MAESTRO was run in order to obtain the resulting modeled network.

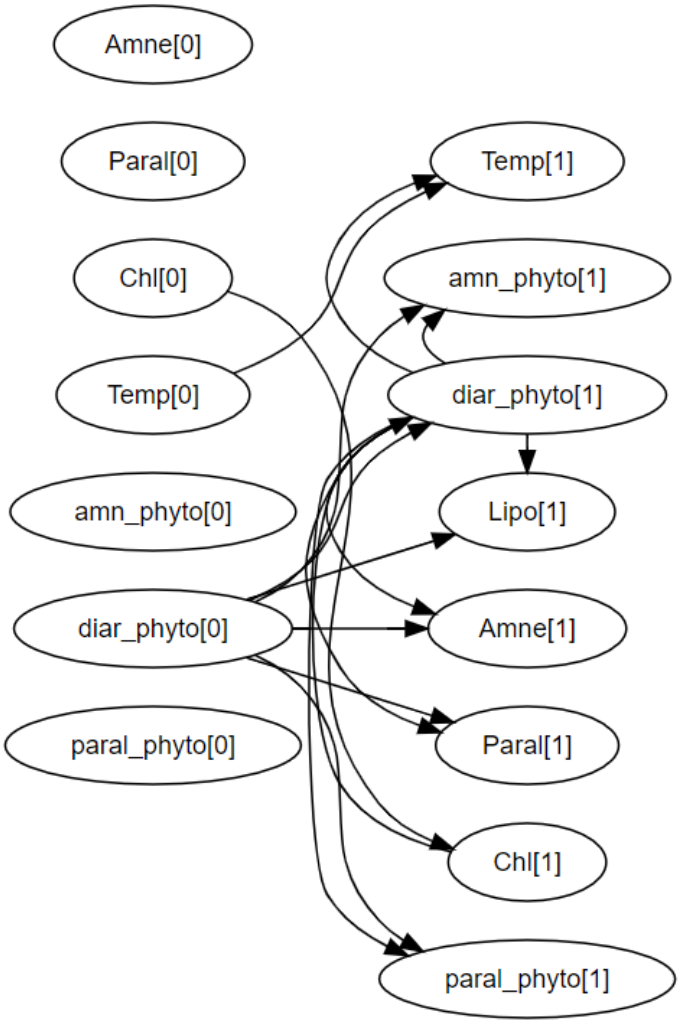

Figure 4.18: Resulting DBN model of the RIAV1 time series.

Complementing this model, Appendix E possesses the conditional probability tables for the generated relations between attributes in the dataset (data already discretized into bins a through d, the former representing low values and the latter the higher values). Unfortunately, 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.

4.2.2 RIAV2 Dataset Presentation

After the same procedure of processing done in RIAV1 was applied to RIAV2, the following plots were obtained:

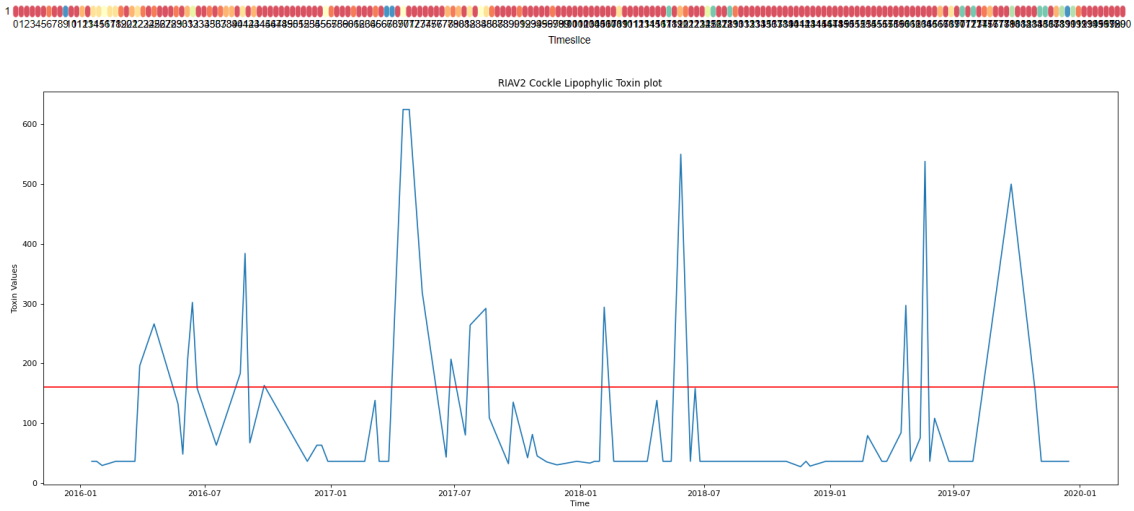

Figure 4.19: Vertical comparison of the lipophilic toxin data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

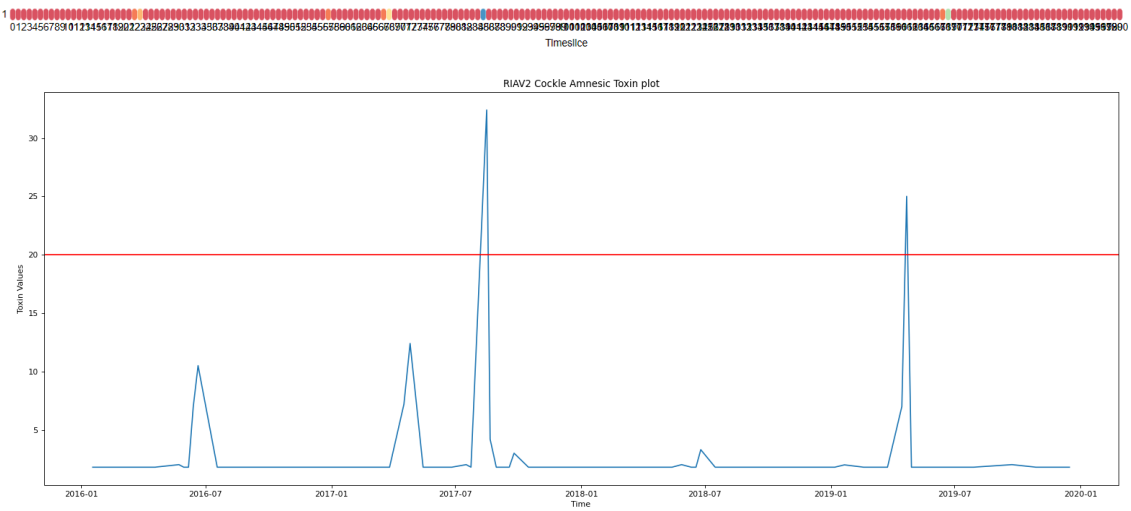

Figure 4.20: Vertical comparison of the amnesic toxin data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

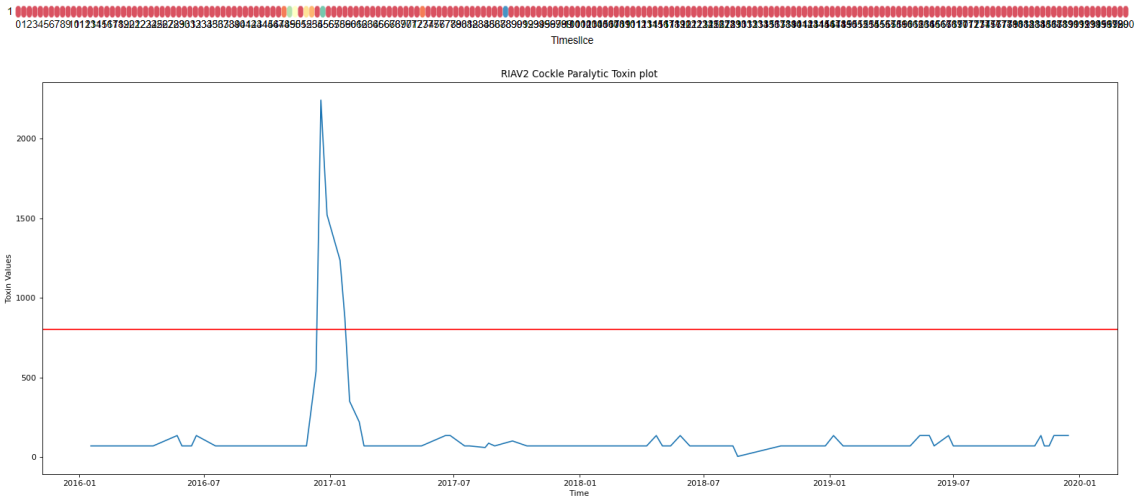

Figure 4.21: Vertical comparison of the paralytic toxin data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

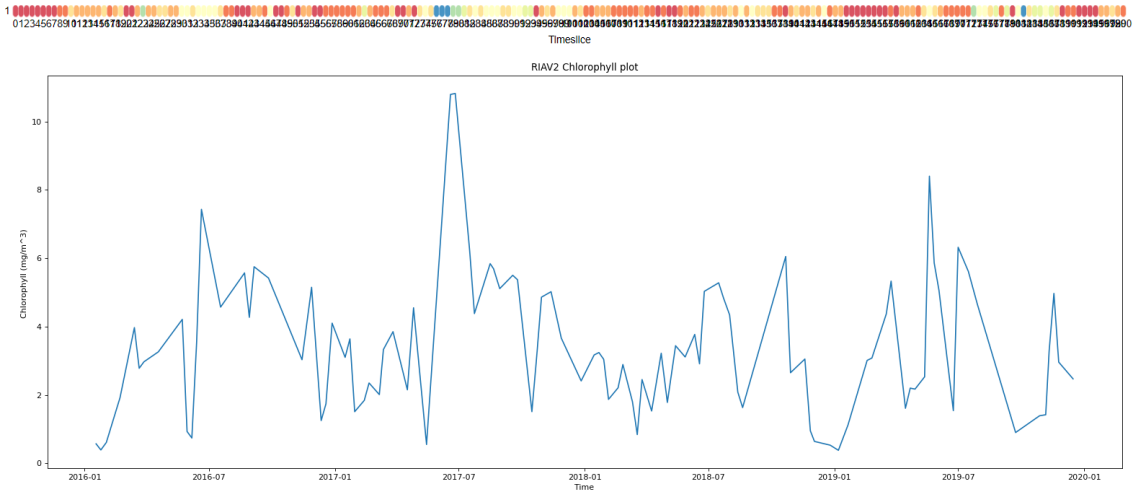

Figure 4.22: Vertical comparison of the chlorophyll data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

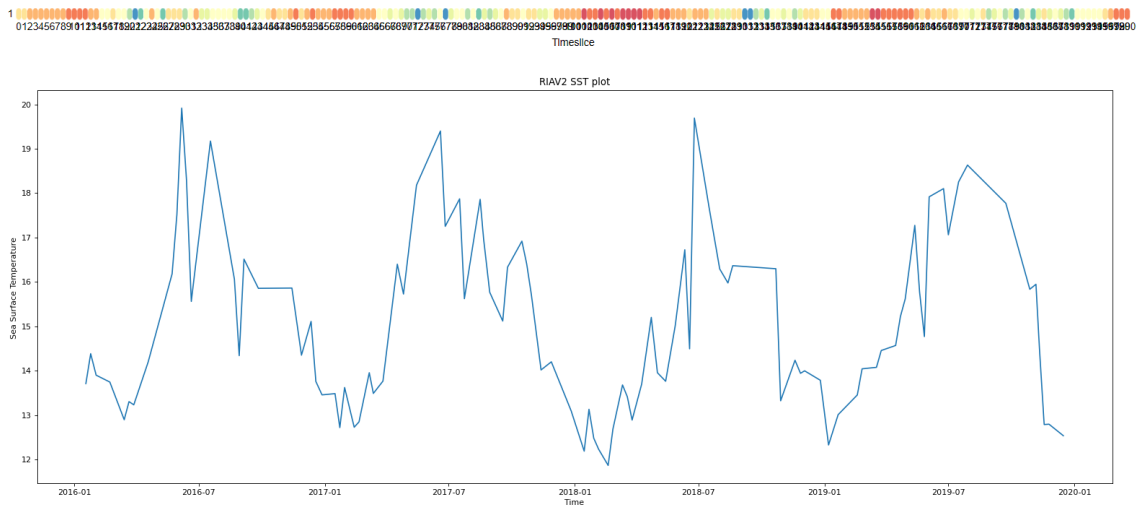

Figure 4.23: Vertical comparison of the sea surface temperature data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

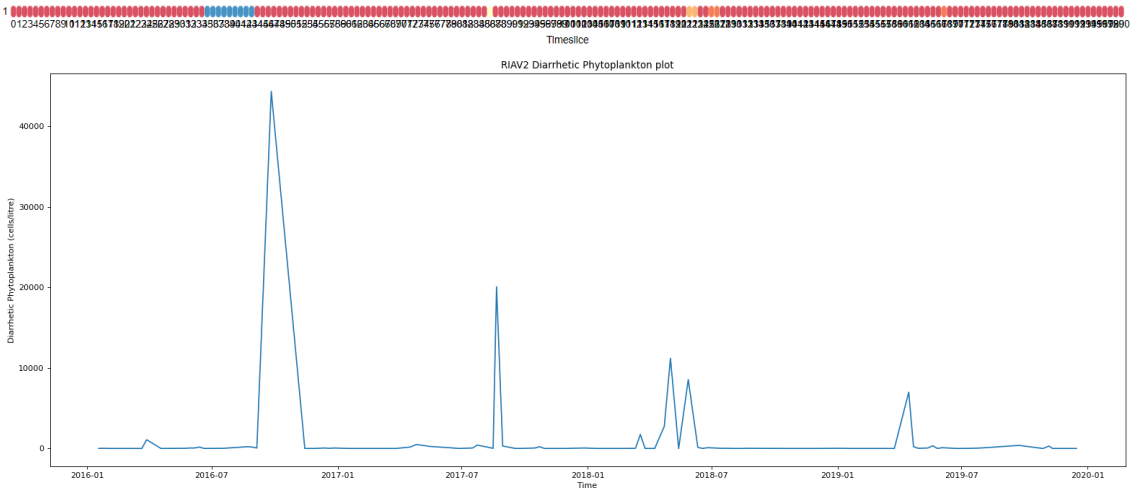

Figure 4.24: Vertical comparison of the DSP producing phytoplankton data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

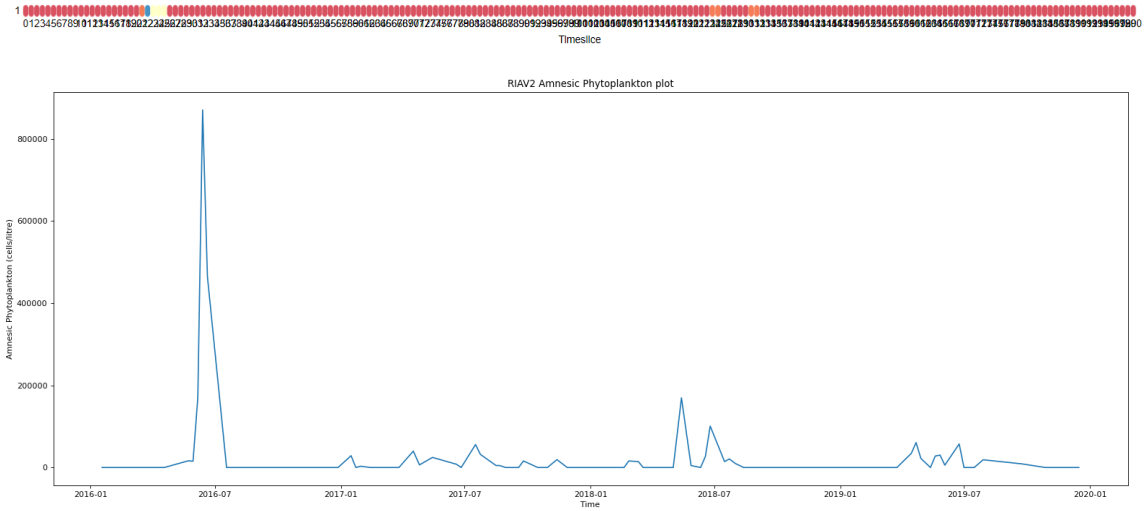

Figure 4.25: Vertical comparison of the ASP producing phytoplankton data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

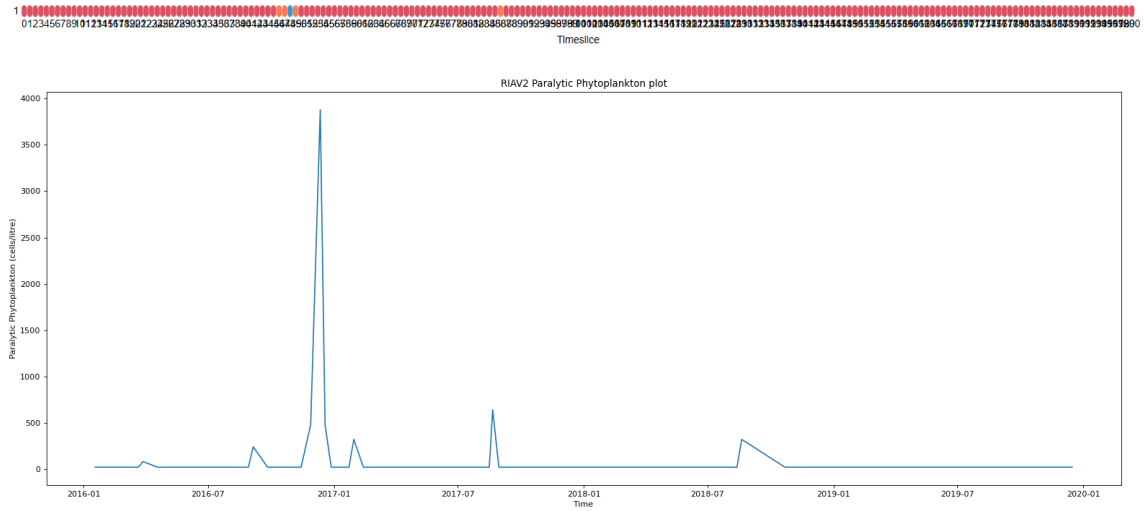

Figure 4.26: Vertical comparison of the PSP producing phytoplankton data of the RIAV2 dataset, MAESTRO (above) and a built plot (below).

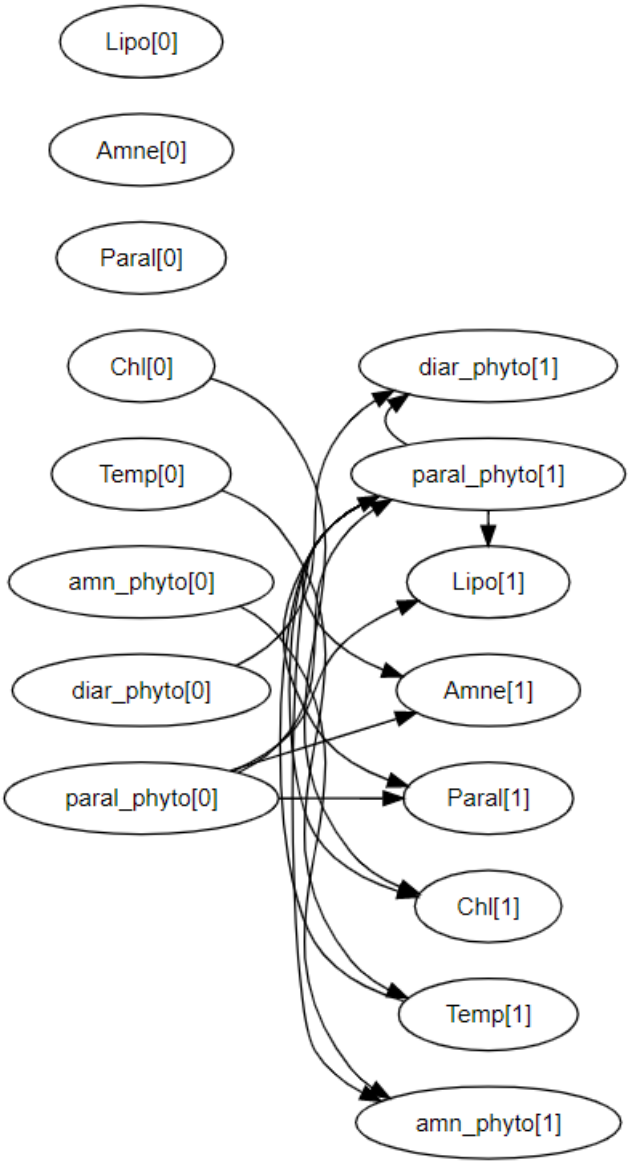

Figure 4.27: Resulting DBN model of the RIAV2 time series.

Referencing Figure 4.13 once more, RIAV2 does not differ much from its RIAV1 counterpart due to an (almost) equally high rate of values assuming the lower threshold value previously defined; however, it is worth mentioning, through observation of the conditional probability tables in Appendix F, that the temperature seemingly has an influence on the PSP producing phytoplankton's concentrations when both are in a higher bin and sea surface temperatures decreases as the PSP producing phytoplankton has a high chance of 50% of lowering their concentration rates (lowering its bin value) compared to higher temperatures).

4.2.3 RIAV3 Dataset Presentation

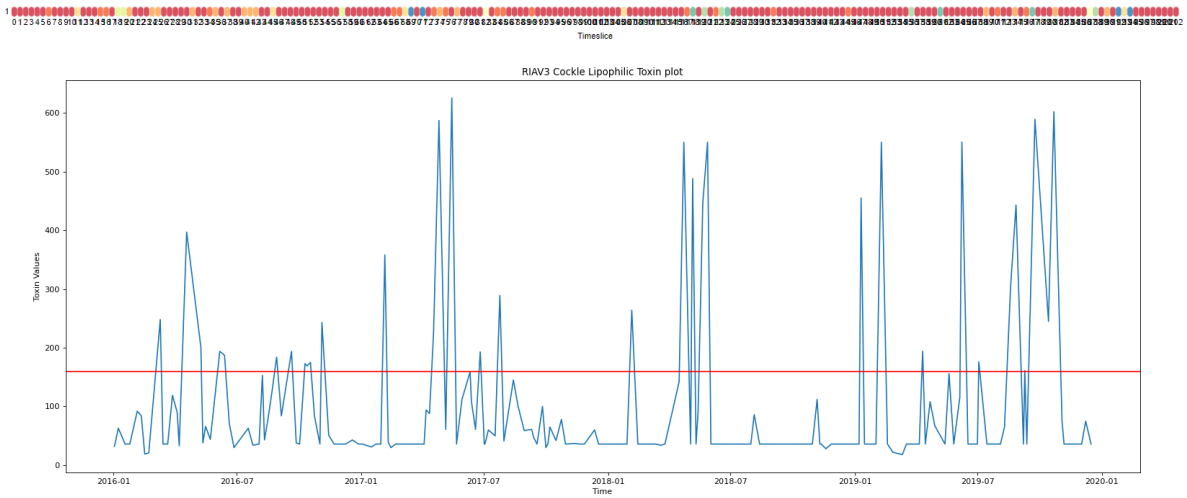

Figure 4.28: Vertical comparison of the lipophilic toxin data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

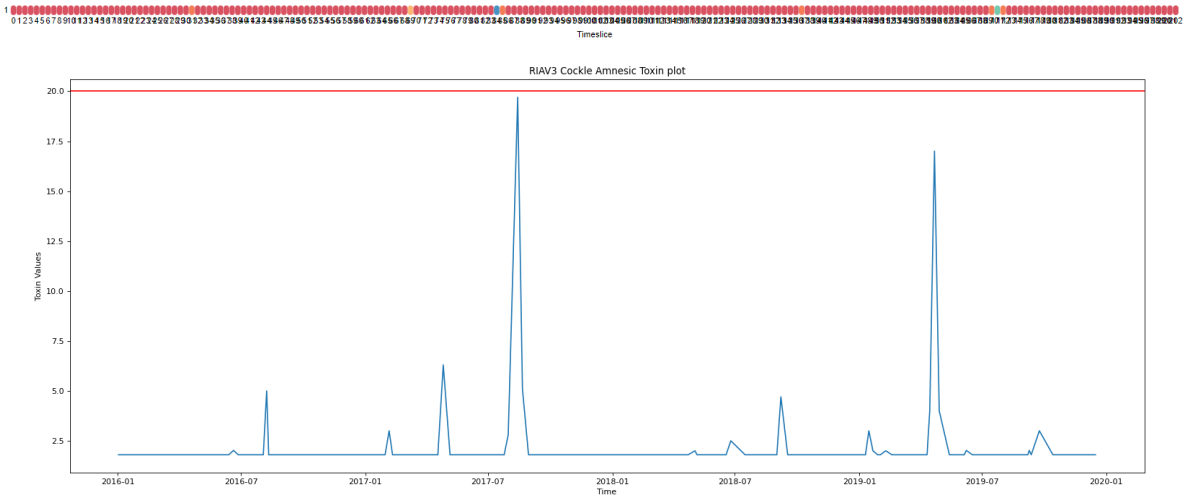

Figure 4.29: Vertical comparison of the amnesic toxin data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

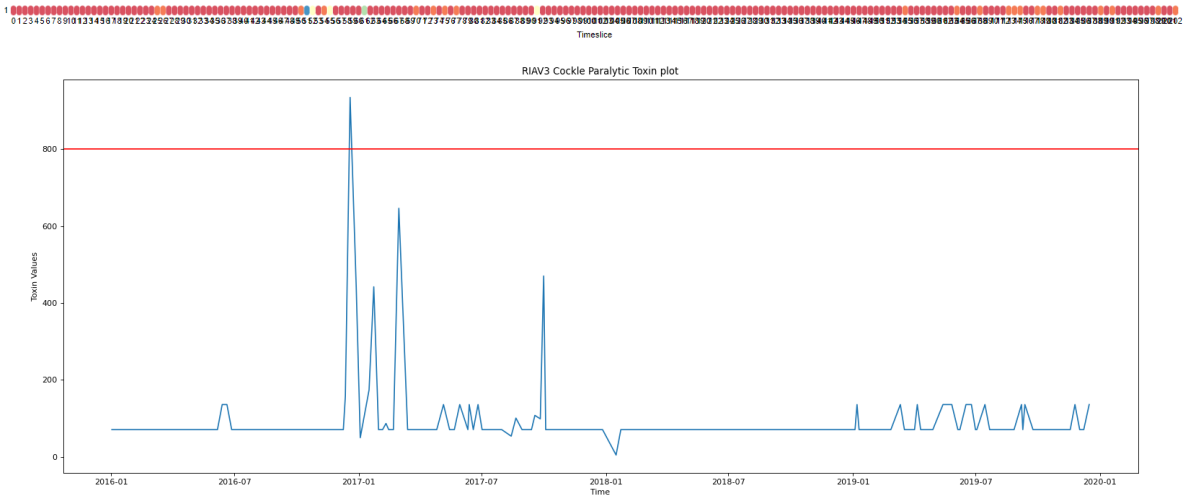

Figure 4.30: Vertical comparison of the paralytic toxin data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

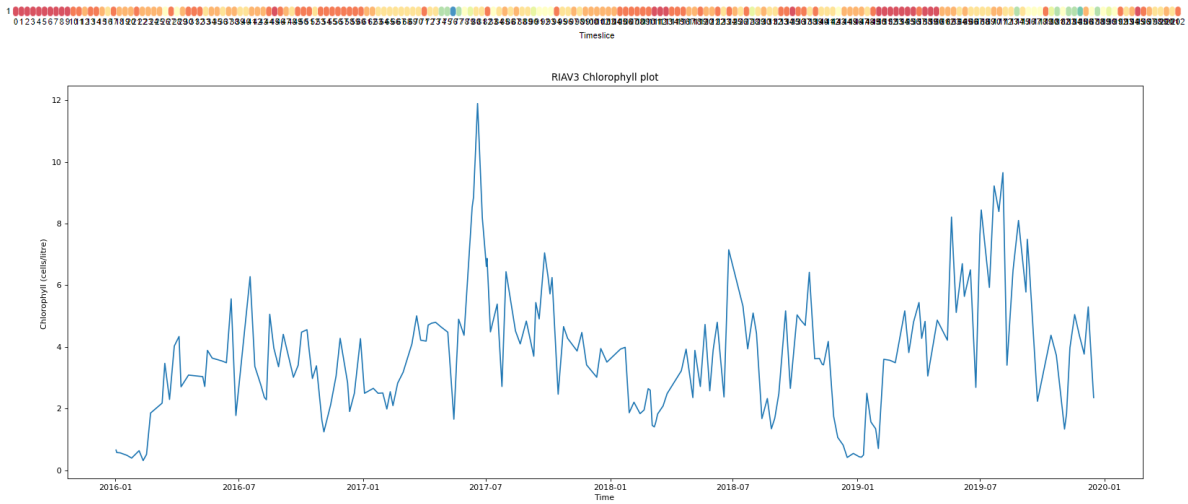

Figure 4.31: Vertical comparison of the chlorophyll data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

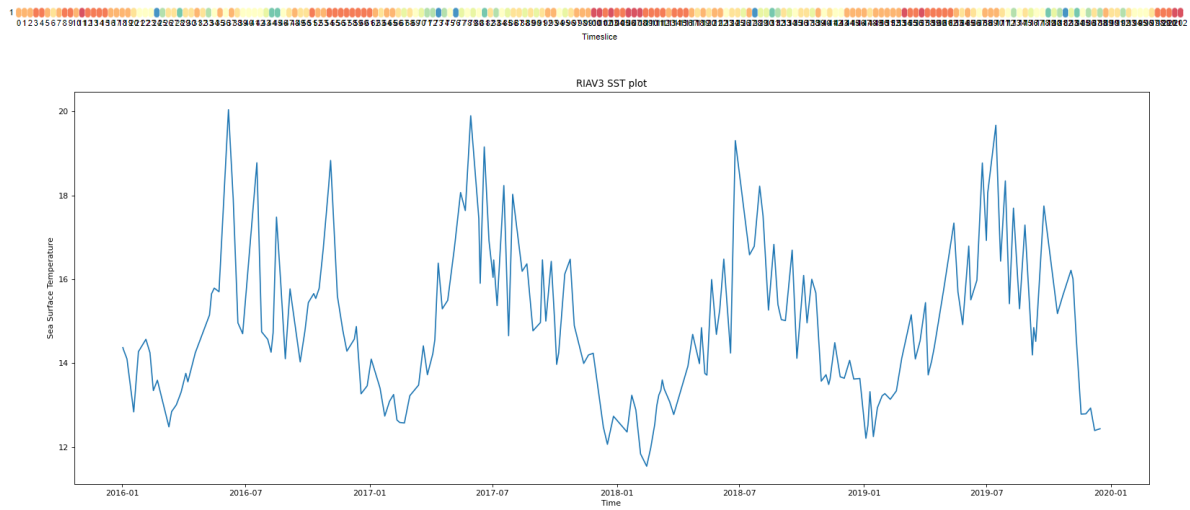

Figure 4.32: Vertical comparison of the sea surface temperature data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

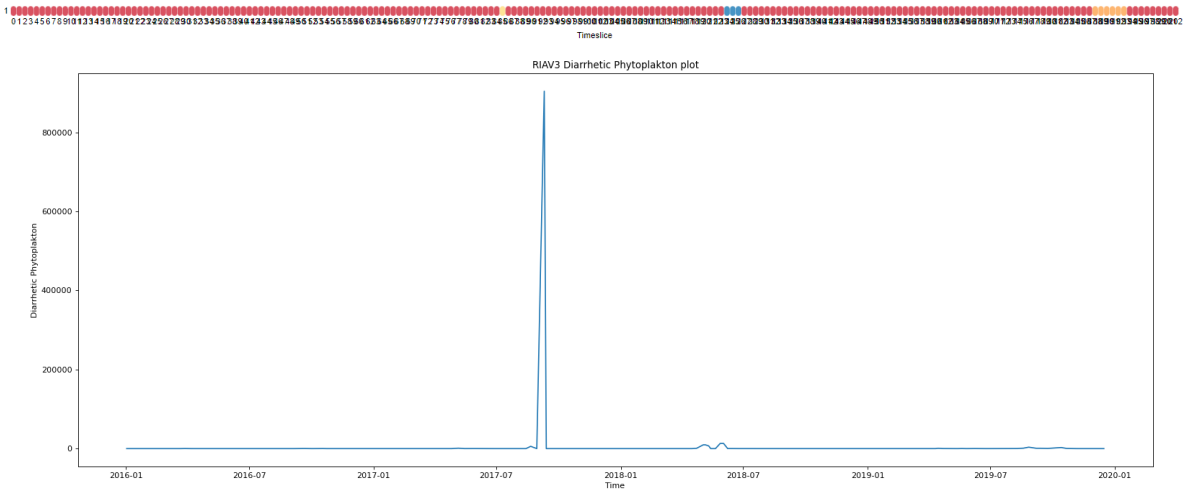

Figure 4.33: Vertical comparison of the DSP producing phytoplankton data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

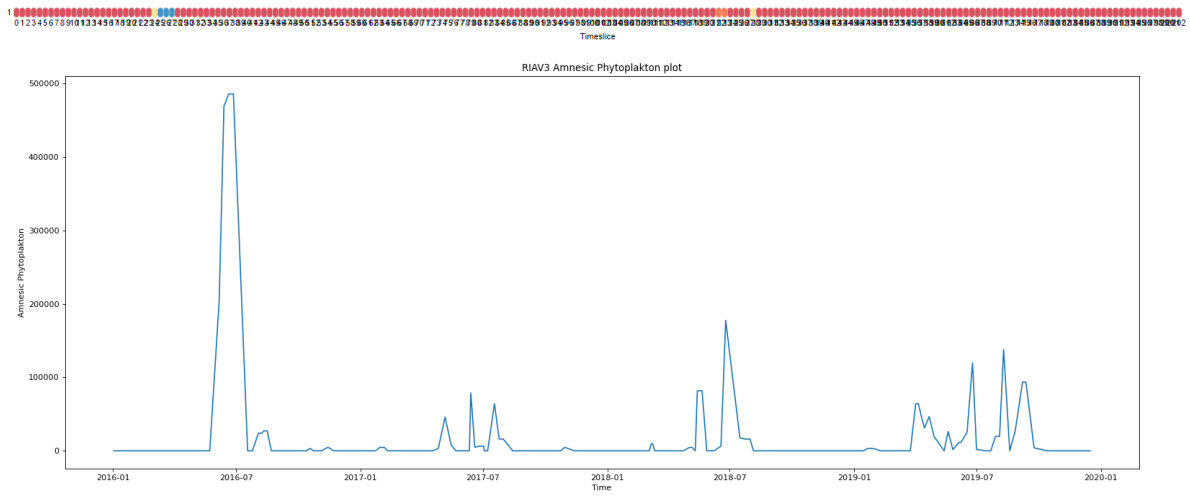

Figure 4.34: Vertical comparison of the ASP producing phytoplankton data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

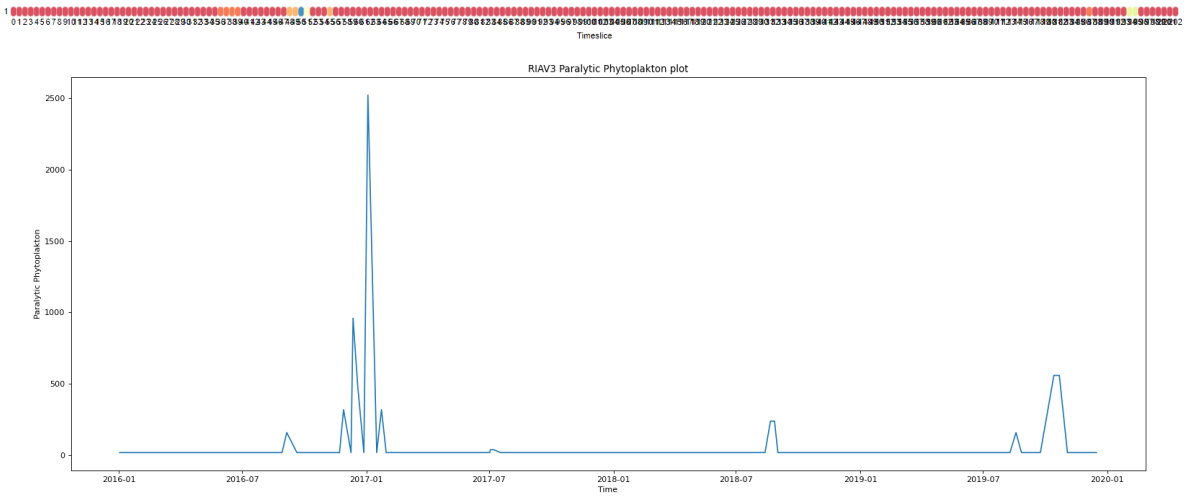

Figure 4.35: Vertical comparison of the PSP producing phytoplankton data of the RIAV3 dataset, MAESTRO (above) and a built plot (below).

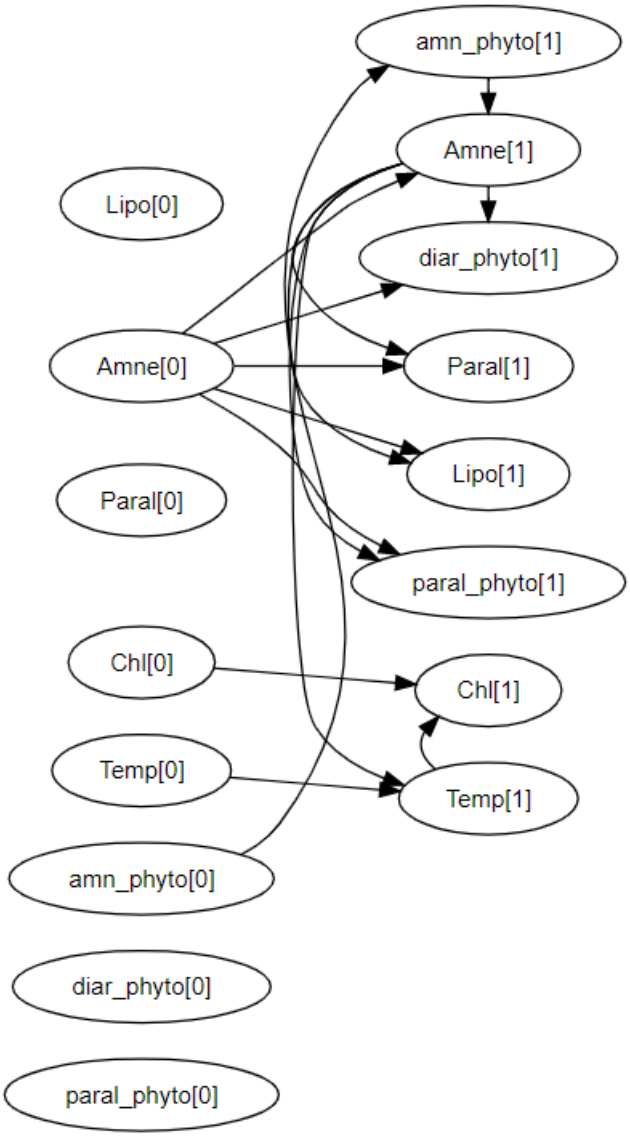

Figure 4.36: Resulting DBN model of the RIAV3 time series.

In the case of the RIAV3 zone (and consulting Appendix G, the SST and Chlorophyll revealed an interesting relation with the amnesic toxin rates detected in the cockle species. Higher values (discretized as c) revealed a higher value of amnesic toxin contaminations detected. Likewise, and taking into consideration the conclusions drawn from observing the RIAV2 DBN model and resulting conditional probability tables, 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).

4.2.4 Combination of RIAV Datasets

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.

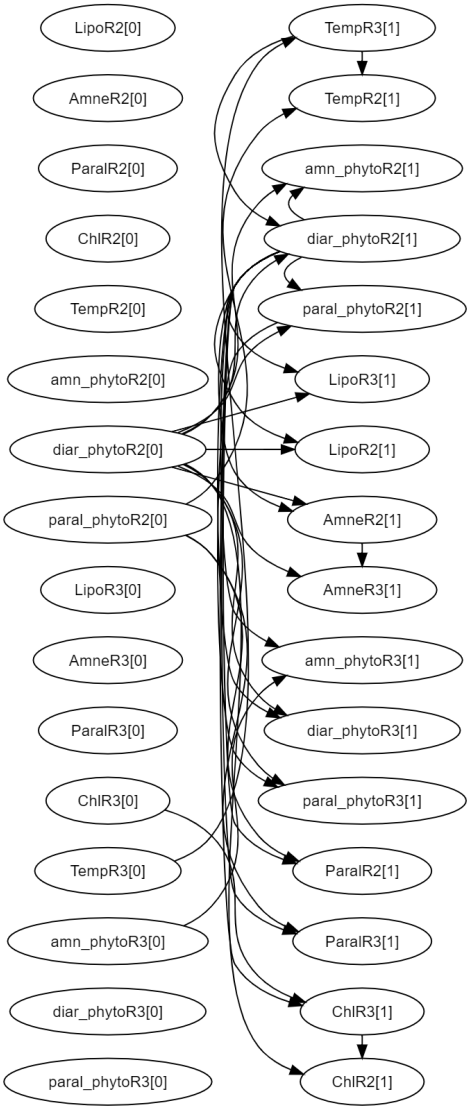

Figure 4.37: MAESTRO's resulting DBN model for the joined time series of RIAV 2 and RIAV3

While a big portion of the results obtained from the DBN model generated (see Figure 4.37) 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, we can observe a probable correlation between chlorophyll, sea surface temperature and biotoxin or phytoplankton concentrations (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.

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

This section will focus on possible improvements for both performance and visualization of the resulting data analysis and forecasting. With the data collected and processed, there are time series with very few data points (see Appendix D) 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. Likewise, the data obtained through Copernicus proved useful in understanding some relations between variables but it's worth noting more attributes could be studied - 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.

Bibliography

[1] "Ipma - legislação e documentação aplicável," 2020, online; 26 December 2020. [Online]. Available: http://www.ipma.pt/pt/bivalves/docs/index.jsp

[2] "Ipma - biotoxinas out19," 2020, online; 26 December 2020. [Online]. Available: http://www.ipma.pt/pt/bivalves/biotox/docs/a-biotoxinas-out19.pdf

[3] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[4] 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*, vol. 53, pp. 33–39, 2016.

[5] 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*, vol. 7, no. 1, pp. 11–25, 2008.

[6] 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*, vol. 114, no. 19, pp. 4975–4980, 2017.

[7] 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.

[8] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress in oceanography*, vol. 123, pp. 123–144, 2014.

[9] 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*, vol. 15, pp. 11–21, 2017.

[10] 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*, vol. 8, no. 4, 2012.

[11] 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.

[12] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull, "Habnet: Machine learning, remote sensing-based detection of harmful algal blooms," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3229–3239, 2020.

[13] J. A. Ekstrom, S. K. Moore, and T. Klinger, "Examining harmful algal blooms through a disaster risk management lens: A case study of the 2015 us west coast domoic acid event," *Harmful algae*, vol. 94, p. 101740, 2020.

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

[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*, vol. 15, no. 7, p. 1322, 2018.

[16] "Copernicus - marine environment monitoring service," 2020, online; 26 December 2020. [Online]. Available: https://marine.copernicus.eu/

[17] 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*, vol. 81, no. 3, pp. 1811–1827, 2016.

[18] 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*, vol. 42, no. 1, pp. 123–139, 2009.

[19] J. Du Preez and S. F. Witt, "Univariate versus multivariate time series forecasting: an application to international tourism demand," *International Journal of Forecasting*, vol. 19, no. 3, pp. 435–451, 2003.

[20] R. H. Shumway and D. S. Stoffer, "Time series analysis and its applications (springer texts in statistics)," 2005.

[21] P. H. Franses, “Seasonality, non-stationarity and the forecasting of monthly time series,” *International Journal of forecasting*, vol. 7, no. 2, pp. 199–208, 1991.

[22] R. Manuca and R. Savit, “Stationarity and nonstationarity in time series analysis,” *Physica D: Nonlinear Phenomena*, vol. 99, no. 2-3, pp. 134–161, 1996.

[23] K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong, “Time series regression studies in environmental epidemiology,” *International journal of epidemiology*, vol. 42, no. 4, pp. 1187–1195, 2013.

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

[25] Y. Yang, “Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation,” *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[26] 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*, vol. 17, no. 2, p. 228, 2012.

[27] H. D.-G. Acquah, “Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship,” *Journal of Development and Agricultural Economics*, vol. 2, no. 1, pp. 001–006, 2010.

[28] K. E. Markon and R. F. Krueger, “An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models,” *Behavior genetics*, vol. 34, no. 6, pp. 593–610, 2004.

[29] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance,” *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[30] G. J. Hahn, “Fitting regression models with no intercept term,” *Journal of Quality Technology*, vol. 9, no. 2, pp. 56–61, 1977.

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

[32] D. R. Osborn and J. P. Smith, “The performance of periodic autoregressive models in forecasting seasonal uk consumption,” *Journal of Business & Economic Statistics*, vol. 7, no. 1, pp. 117–127, 1989.

[33] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[34] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] A. Liaw, M. Wiener *et al.*, “Classification and regression by randomforest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[36] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” in *International workshop on machine learning and data mining in pattern recognition*. Springer, 2012, pp. 154–168.

[37] 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*, vol. 11, no. 1, pp. 1–11, 2021.

[38] S. Basu, K. Kumbier, J. B. Brown, and B. Yu, “Iterative random forests to discover predictive and stable high-order interactions,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1943–1948, 2018.

[39] 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*, vol. 9, no. 1, pp. 1–9, 2019.

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

[41] D. J. Hill, B. S. Minsker, and E. Amir, “Real-time bayesian anomaly detection for environmental sensor data,” in *Proceedings of the Congress-International Association for Hydraulic Research*, vol. 32, no. 2. Citeseer, 2007, p. 503.

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

[43] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden markov model.” in *CVPR*, vol. 92, 1992, pp. 379–385.

[44] A. Palmer, J. J. Montano, and A. Sesé, “Designing an artificial neural network for forecasting tourism time series,” *Tourism management*, vol. 27, no. 5, pp. 781–790, 2006.

[45] F. Recknagel, M. French, P. Harkonen, and K.-I. Yabunaka, “Artificial neural network approach for modelling and prediction of algal blooms,” *Ecological Modelling*, vol. 96, no. 1-3, pp. 11–28, 1997.

[46] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[47] 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*, 2016, pp. 785–794.

[48] M. Gumus and M. S. Kiran, “Crude oil price forecasting using xgboost,” in *2017 International conference on computer science and engineering (UBMK)*. IEEE, 2017, pp. 1100–1103.

[49] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, “Machine learning–xgboost analysis of language networks to classify patients with epilepsy,” *Brain informatics*, vol. 4, no. 3, pp. 159–169, 2017.

A

Zone information and respective geographical coordinates

Zona do país	Zona de produção	Código	Local de Amostragem	Latitude	Longitude
Norte	Estuário do Lima	ELM	Montante da Ponte Eiffel	41,69164	-8,81375
	Litoral Viana	L1	Carreço	41,74283	-8,87833
			Labruge	41,27802	-8,76900
	Litoral Matosinhos	L2	Aguda	41,03335	-8,70167
			Leça da Palmeira	41,19667	-8,71114
Centro	Ria de Aveiro, Triângulo das Correntes/Moacha	RIAV1	Moacha	40,69556	-8,71167
			Piscicultura	40,68663	-8,71390
	Ria de Aveiro, Canal de Mira	RIAV2	Costa Nova	40,61720	-8,74235
			Ponte da Barra	40,63020	-8,73822
			Sul da Ponte da Barra	40,62160	-8,73934
	Ria de Aveiro, Canal Principal Espinheiro	RIAV3	Canal do Espinheiro	40,67667	-8,68444
			Ilha dos Puxadoiros	40,65444	-8,66833
Centro	Ria de Aveiro, Canal de Ílhavo	RIAV4	Corte das Freiras	40,63002	-8,67679
			Sul da Ponte da A25	40,62965	-8,68567
	Litoral Aveiro	L3	Torreira	40,75883	-8,80100
	Estuário do Mondego – Braço	EMN1	Morraceira Norte	40,14013	-8,82693
	Estuário do Mondego – Braço Sul	EMN2	Morraceira Sul	40,11936	-8,83060
	Litoral Fig. da Foz – Nazaré	L4	Leirosa	40,05605	-8,89238
			Pedrógão	39,91845	-8,99620
Lisboa e Vale do Tejo	Lagoa de Óbidos	LOB	Espichel	39,40392	-9,20923
			Greijau	39,42000	-9,21245
	Estuário do Tejo	ETJ	Baliza de Ferro [2)]	38,77817	-9,03867
			Trafaria [1)]	38,67333	-9,25917
			Cacilhas	38,68944	-9,14091
			Samouco	38,76259	-9,00108
			Alcochete	38,77303	-9,00108
	Litoral Peniche – Cabo Raso	L5a	Praia dos Coxos	38,00468	-9,42503

Lisboa e Vale do Tejo	Litoral Cabo Raso – Lagoa de Albufeira	L5b	Praia da Rainha	38,614473	-9,22468
			Praia do Norte	38,63790	-9,24342
			Costa de Caparica	38,64562	-9,24341
Alentejo	Lagoa de Albufeira	LAL	Jangada	38,51254	-9,17211
			Lagoa	38,51308	-9,17554
	Estuário do Sado – Esteiro da Marateca	ESD1	Faralhão	38,52540	-8,79866
			Canal da Vaia	38,54224	-8,79039
				38,54260	-8,79201
			Mitrena	38,50800	-8,81199
	Estuário do Sado – Canal de Alcácer	ESD2	Abul	38,42794	-8,68342
			Palma	38,40878	-8,64487
			Carrasqueira	38,42680	-8,71925
	Estuário do Mira	EMR	Troviscais	37,67259	-8,7254
			Jusante da Ponte	37,72790	-8,77173
	Litoral Setúbal-Sines	L6	Sines	38,08975	-8,82856
			Comporta	38,42633	-8,84762
			Praia da Costa do [illegible]	37,96875	-8,87341
Algarve	Ria de Alvor	LAG	Vale da Lama	37,13400	-8,62205
				37,12475	-8,62893
	Rio Arade – Montante da Ponte [illegible]	POR1	Rio Arade	37,15733	-8,50235
	Ria de Alvor – Povoação	POR2	Povoação	37,13217	-8,59750
	Rio Arade – Parchal	POR3	Parchal	37,13861	-8,51218
	Litoral Aljezur – S. Vicente	L7a	Aljezur – Amoreira	37,35505	-8,84654
			Aljezur	37,29571	-8,87083
	Litoral *Offshore*	L7b	Sagres – Cultura	37,02250	-8,88583
	Litoral S. Vicente – Lagos	L7c1	*Offshore*	37,01666	-8,86920
			Ponta do Zavial	37,03423	-8,84875
	Litoral Lagos – Albufeira	L7c2	Porto de Mós	37,06588	-8,68550
			Albufeira	37,08257	-8,18077
	Ria Formosa – Faro-Cais Novo – Geada	FAR1	Marchil	37,01867	-7,94833
				37,00200	-7,94833
	Ria Formosa – Faro- Regato de Azeites – Barrinha	FAR2	Largura	36,99667	-7,96597

Algarve	Ria Formosa – Olhão	OLH1	Regueira de Água Quente	37,03452	-7,78813
		OLH2	Fortaleza	37,02550	-7,81250
		OLH3	Ilhote Negro [b)]	37,01561	-7,85518
		OLH4	Garganta	37,00517	-7,86867
		OLH5	Culatra	36,99333	-7,84667
				36,99450	-7,86100
	Ria Formosa – Fuzeta	FUZ	Fuzeta	37,02362	-7,44591
	Litoral Faro-Olhão	L8	Culatra	36,98550	-7,83383
	Ria Formosa – Tavira	TAV	Quatro Águas	37,11385	-7,63050
	Ria Formosa – Cacela	VT	Cacela	37,15385	-7,55191
	Rio Guadiana	GUA	Castro Marim	37,21413	-7,43194
Algarve	Litoral Tavira – V.R.S.A.	L9	Monte Gordo	37,17500	-7,43733

B

Zones and respective species captured

Zona do país	Zona de produção	Código	Local de Amostragem	Espécies Amostradas
Norte	Estuário do Lima	ELM	Montante da Ponte Eiffel	Ostra-portuguesa
	Litoral Viana	L1	Carreço	Mexilhão
				Ouriço-do-mar
			Labruge	Amêijoa-branca
	Litoral Matosinhos	L2	Aguda	Amêijoa-branca
				Castanhola
			Leça da Palmeira	Ouriço-do-mar
Centro	Ria de Aveiro, Triângulo das Correntes/Moacha	RIAV1	Moacha	Amêijoa-macha
				Berbigão
				Longueirão
			Piscicultura	Ostra-japonesa ou gigante
	Ria de Aveiro, Canal de Mira	RIAV2	Costa Nova	Ostra-japonesa ou gigante
			Ponte da Barra	Amêijoa-macha
			Sul da Ponte da Barra	Berbigão
				Longueirão
	Ria de Aveiro, Canal Principal Espinheiro	RIAV3	Canal do Espinheiro	Amêijoa-macha
				Berbigão
				Longueirão
			Ilha dos Puxadoiros	Ostra-japonesa ou gigante
Centro	Ria de Aveiro, Canal de Ílhavo	RIAV4	Corte das Freiras	Ostra-japonesa ou gigante
			Sul da Ponte da A25	Berbigão
				Longueirão
				Amêijoa-macha
	Litoral Aveiro	L3	Torreira	Amêijoa-branca
	Estuário do Mondego – Braço	EMN1	Morraceira Norte	Berbigão
	Estuário do Mondego – Braço Sul	EMN2	Morraceira Sul	Berbigão
				Lambujinha
	Litoral Fig. da Foz – Nazaré	L4	Leirosa	Mexilhão
			Pedrógão	Amêijoa-branca
Lisboa e Vale do Tejo	Lagoa de Óbidos	LOB	Espichel	Amêijoa-boa
				Amêijoa-japonesa
				Berbigão
			Greijau	Amêijoa-macha
	Estuário do Tejo	ETJ	Baliza de Ferro [2)]	Amêijoa-japonesa
			Trafaria [1)]	Amêijoa-macha
			Cacilhas	Pé-de-burro
			Samouco	Amêijoa-japonesa
				Lambujinha
			Alcochete	Amêijoa-japonesa
	Litoral Peniche – Cabo Raso	L5a	Praia dos Coxos	Mexilhão
				Ouriço-do-mar

Lisboa e Vale do Tejo	Litoral Cabo Raso – Lagoa de Albufeira	L5b	Praia da Rainha	Conquilha
			Praia do Norte	Longueirão
			Costa de Caparica	Mexilhão
Alentejo	Lagoa de Albufeira	LAL	Jangada	Mexilhão
			Lagoa	Berbigão
				Amêijoa-boa
	Estuário do Sado – Esteiro da Marateca	ESD1	Faralhão	Amêijoa-japonesa
				Berbigão
			Canal da Vaia	Lambujinha
				Ostra-portuguesa
			Mitrena	Ostra-plana
	Estuário do Sado – Canal de Alcácer	ESD2	Abul	Ostra-portuguesa
			Palma	Lambujinha
			Carrasqueira	Amêijoa-japonesa
				Longueirão
	Estuário do Mira	EMR	Troviscais	Ostra-portuguesa
			Jusante da Ponte	Mexilhão
	Litoral Setúbal-Sines	L6	Sines	Amêijoa-branca
			Comporta	Ameijola
				Conquilha
				Longueirão
			Praia da Costa do	Ouriço-do-mar
Algarve	Ria de Alvor	LAG	Vale da Lama	Amêijoa-boa
				Ostra-japonesa ou gigante
	Rio Arade – Montante da Ponte	POR1	Rio Arade	Amêijoa-boa
	Ria de Alvor – Povoação	POR2	Povoação	Amêijoa-boa
				Mexilhão
	Rio Arade – Parchal	POR3	Parchal	Ostra-japonesa ou gigante
	Litoral Aljezur – S. Vicente	L7a	Aljezur – Amoreira	Mexilhão
			Aljezur	Ouriço-do-mar
	Litoral *Offshore*	L7b	Sagres – Cultura	Ostra-japonesa ou gigante
	Litoral S. Vicente – Lagos	L7c1	*Offshore*	Mexilhão
			Ponta do Zavial	Ostra-japonesa ou gigante
	Litoral Lagos – Albufeira	L7c2	Porto de Mós	Mexilhão
			Albufeira	Amêijoa-branca
				Conquilha
				Pé-de-burrinho
	Ria Formosa – Faro-Cais Novo – Geada	FAR1	Marchil	Amêijoa-boa
				Berbigão
	Ria Formosa – Faro- Regato de Azeites – Barrinha	FAR2	Largura	Amêijoa-boa
				Berbigão
				Ostra-japonesa ou gigante

Algarve	Ria Formosa – Olhão	OLH1	Regueira de Água Quente	Amêijoa-boa
				Ostra-japonesa ou gigante
		OLH2	Fortaleza	Amêijoa-boa
				Amêijoa-cão
				Ostra-japonesa ou gigante
		OLH3	Ilhote Negro [b)]	Amêijoa-boa
		OLH4	Garganta	Amêijoa-boa
				Ostra-japonesa ou gigante
		OLH5	Culatra	Berbigão
				Ostra-japonesa ou gigante
				Amêijoa-boa
	Ria Formosa – Fuzeta	FUZ	Fuzeta	Amêijoa-boa
				Berbigão
				Ostra-japonesa ou gigante
	Litoral Faro-Olhão	L8	Culatra	Amêijoa-branca
				Conquilha
				Pé-de-burrinho
	Ria Formosa – Tavira	TAV	Quatro Águas	Amêijoa-boa
				Mexilhão
				Ostra-japonesa ou gigante
	Ria Formosa – Cacela	VT	Cacela	Ostra-japonesa ou gigante
	Rio Guadiana	GUA	Castro Marim	Ostra-japonesa ou gigante
Algarve	Litoral Tavira – V.R.S.A.	L9	Monte Gordo	Amêijoa-branca
				Conquilha
				Pé-de-burrinho

C

Zones evolution since the start of data logging

	2014	2015	2016	2017	2018	2019
Norte	ELM	ELM	ELM	ELM	ELM	ELM
	L1	L1	L1	L1	L1	L1
	L2	L2	L2	L2	L2	L2
Centro	L3	L3	L3	L3	L3	L3
	L4	L4	L4	L4	L4	L4
	EMN1	EMN1	EMN1	EMN1	EMN1	EMN1
	EMN2	EMN2	EMN2	EMN2	EMN2	EMN2
	RIAV1	RIAV1	RIAV1	RIAV1	RIAV1	RIAV1
	RIAV2	RIAV2	RIAV2	RIAV2	RIAV2	RIAV2
	RIAV3	RIAV3	RIAV3	RIAV3	RIAV3	RIAV3
	RIAV4	RIAV4	RIAV4	RIAV4	RIAV4	RIAV4
Lisboa e Vale do Tejo	L5	L6	L7	L5a	L5a	L5a
	ETJ	ETJ	ETJ	L5b	L5b	L5b
	LOB	LOB	LOB	ETJ	ETJ	ETJ
	-	-	-	LOB	LOB	LOB
Alentejo	L6	L6	L6	L6	L6	L6
	EMR	EMR	EMR	EMR	EMR	EMR
	ESD1	ESD1	ESD1	ESD1	ESD1	ESD1
	ESD2	ESD2	ESD2	ESD2	ESD2	ESD2
	LAL	LAL	LAL	LAL	LAL	LAL
Algarve	L7a	L7a	L7a	L7a	L7a	L7a
	L7b	L7b	L7b	L7b	L7b	L7b
	L7c	L7c	L7c	L7c	L7c1	L7c1
	-	-	-	-	L7c2	L7c2
	L8	L8	L8	L8	L8	L8
	L9	L9	L9	L9	L9	L9
	FAR1	FAR1	FAR1	FAR1	FAR1	FAR1
	FAR2	FAR2	FAR2	FAR2	FAR2	FAR2
	FUZ1	FUZ2	FUZ3	FUZ4	FUZ5	FUZ6
	-	-	-	GUA	GUA	GUA
	LAG	LAG	LAG	LAG	LAG	LAG
	OLH1	OLH1	OLH1	OLH1	OLH1	OLH1
	OLH2	OLH2	OLH2	OLH2	OLH2	OLH2
	OLH3	OLH3	OLH3	OLH3	OLH3	OLH3
	OLH4	OLH4	OLH4	OLH4	OLH4	OLH4
	OLH5	OLH5	OLH5	OLH5	OLH5	OLH5
	POR2	POR2	POR2	POR2	POR2	POR2
	-	-	POR3	POR3	POR3	POR3
	TAV2	TAV2	TAV2	TAV2	TAV	TAV
	VT1	VT1	VT1	VT1	VT	VT

D

Sample count of each species in each region

		Ostra-Portuguesa	Mexilhão	Amêijoa-branca	Amêijoa-boa	Amêijoa-macha	Castanhola
Norte	ELM	c = 20	x	x	x	x	x
	L1	x	c = 137	c=27	x	x	x
	L2	x	c =156	c =156	x	x	c=32
Centro	RIAV1	x	x	x	x	c=45	x
	RIAV2	x	x	x	x	c=45	x
	RIAV3	x	x	x	x	c=45	x
	RIAV4	x	x	x	x	c=13	x
	L3	x	x	c=83	x	x	x
	EMN1	x	x	x	x	x	x
	EMN2	x	x	x	x	x	x
	L4	x	c=62	c=43	c=97	x	x
Lisboa e Vale do Tejo	LOB	x	x	x	c=9	x	x
	ETJ	x	x	x	x	c=9	x
	L5a	x	c=77	x	x	x	x
	L5b	x	c=138	x	x	x	x
Alentejo	LAL	x	c=189	x	c=28	x	x
	ESD1	c=1	x	x	x	x	x
	ESD2	x	x	x	x	x	x
	EMR	c=18	c=73	x	x	x	x
	L6	x	x	c=54	x	x	x
Algarve	LAG	x	x	x	c=15	x	x
	POR1	x	x	x	x	x	x
	POR2	x	c=152	x	x	x	x
	POR3	x	x	x	x	x	x
	L7a	x	c=94	x	x	x	x
	L7b	x	x	x	x	x	x
	L7c1	x	c=174	x	x	x	x
	L7c2	x	c=160	c=2	x	x	x
	FAR1	x	x	x	c=11	x	x
	FAR2	x	c=55	x	x	x	x
	OLH1	x	x	x	c=69	x	x
	OLH2	x	x	x	c=5	x	x
	OLH3	x	x	x	c=5	x	x
	OLH4	x	x	x	c=4	x	x
	OLH5	x	x	x	c=9	x	x
	FUZ	x	x	x	c=97	x	x
	L8	x	x	c=24	x	x	x
	TAV	x	c=176	x	c=19	x	x
	VT	x	x	x	x	x	x
	GUA	x	x	x	x	x	x
	L9	x	x	x	x	x	x

		Berbigão	Longueirão	Ostra-japonesa/gigante	Lambujinha	Pé-de-burro	Conquilha
Norte	ELM	x	x	x	x	x	x
	L1	x	x	x	x	x	x
	L2	x	x	x	x	x	x
Centro	RIAV1	c=206	c=42	c=26	x	x	x
	RIAV2	c=202	c=64	c=25	x	x	x
	RIAV3	c=199	c=11	c=20	x	x	x
	RIAV4	c=29	c=16	c=13	x	x	x
	L3	x	x	x	x	x	x
	EMN1	c=79	x	x	x	x	x
	EMN2	c=35	x	x	c=49	x	x
	L4	x	x	x	x	x	x
Lisboa e Vale do Tejo	LOB	c=191	x	x	c=1	x	x
	ETJ	x	x	x	x	x	x
	L5a	x	x	x	x	x	x
	L5b	x	c=37	x	x	x	c=53
Alentejo	LAL	c=27	x	x	x	x	x
	ESD1	c=8	x	x	c=93	x	x
	ESD2	x	c=3	x	c=29	x	x
	EMR	x	x	x	x	x	x
	L6	x	x	x	x	x	x
Algarve	LAG	x	c=75	c=11	x	x	c=92
	POR1	x	x	x	x	x	x
	POR2	x	x	x	x	x	x
	POR3	x	x	c=12	x	x	x
	L7a	x	x	x	x	x	x
	L7b	x	x	c=97	x	x	x
	L7c1	x	x	c=1	x	x	x
	L7c2	x	x	x	x	x	c=8
	FAR1	c=7	x	x	x	x	x
	FAR2	c=98	x	c=8	x	x	x
	OLH1	x	x	c=5	x	x	x
	OLH2	x	x	c=6	x	x	x
	OLH3	x	x	x	x	x	x
	OLH4	x	x	c=3	x	x	x
	OLH5	c=112	x	c=11	x	x	x
	FUZ	c=18	x	c=8	x	x	x
	L8	x	x	x	x	x	c=166
	TAV	x	x	c=13	x	x	x
	VT	x	x	c=95	x	x	x
	GUA	x	x	c=124	x	x	x
	L9	x	x	x	x	x	c=188

		Ostra-plana	Ameijola	Amêijoa-japonesa	Amêijoa-Cão
Norte	ELM	x	x	x	x
	L1	x	x	x	x
	L2	x	x	x	x
Centro	RIAV1	x	x	x	x
	RIAV2	x	x	x	x
	RIAV3	x	x	x	x
	RIAV4	x	x	x	x
	L3	x	x	x	x
	EMN1	x	x	x	x
	EMN2	x	x	x	x
	L4	x	x	x	x
Lisboa e Vale do Tejo	LOB	x	x	c=10	x
	ETJ	x	x	x	x
	L5a	x	x	x	x
	L5b	x	x	x	x
Alentejo	LAL	x	x	x	x
	ESD1	c=1	x	c=2	x
	ESD2	x	x	c=102	x
	EMR	x	x	x	x
	L6	x	c=56	x	x
Algarve	LAG	x	x	x	x
	POR1	x	x	x	x
	POR2	x	x	x	x
	POR3	x	x	x	x
	L7a	x	x	x	x
	L7b	x	x	x	x
	L7c1	x	x	x	x
	L7c2	x	x	x	x
	FAR1	x	x	x	x
	FAR2	x	x	x	x
	OLH1	x	x	x	x
	OLH2	x	x	x	c=5
	OLH3	x	x	x	x
	OLH4	x	x	x	x
	OLH5	x	x	x	x
	FUZ	x	x	x	x
	L8	x	x	x	x
	TAV	x	x	x	x
	VT	x	x	x	x
	GUA	x	x	x	x
	L9	x	x	x	x

E

MAESTRO generated conditional probability tables for RIAV 1 zone

DSP Phyto (lag = 1)	DSP Phyto(lag = 0)	P(Lipophilic Toxins = a)	P(Lipophilic Toxins = b)	P(Lipophilic Toxins = d)	P(Lipophilic Toxins = c)
a	a	0.931	0.036	0.017	0.017
a	d	1	0	0	0
d	d	1	0	0	0
d	a	1	0	0	0

DSP Phyto (lag = 1)	DSP Phyto(lag = 0)	P(Amnesic Toxins = a)	P(Lipophilic Toxins = d)	P(Lipophilic Toxins = c)
d	a	1	0	0
a	d	1	0	0
d	d	1	0	0
a	a	0.990	0.003	0.007

DSP Phyto (lag = 1)	DSP Phyto(lag = 0)	P(Paralytic Toxins = a)	P(Paralytic Toxins = d)	P(Paralytic Toxins = b)
d	d	1	0	0
d	a	1	0	0
a	a	0.983	0.007	0.010
a	d	1	0	0

Chlorophyll(lag = 1)	P(Chlorophyll = a)	P(Chlorophyll = b)	P(Chlorophyll = c)	P(Chlorophyll = d)
d	0	0.200	0.400	0.400
b	0.149	0.784	0.061	0.007
a	0.795	0.172	0.033	0.000
c	0.094	0.312	0.531	0.062

SST (lag = 1)	DSP Phyto(lag = 0)	P(SST = b)	P(SST = a)	P(SST = c)	P(SST = d)
b	a	0.741	0.086	0.144	0.029
b	d	0.500	0	0.500	0
d	a	0.154	0	0.346	0.500
c	d	1	0	0	0
d	d	0.250	0.250	0.250	0.250
a	d	0.250	0.250	0.250	0.250
a	a	0.133	0.867	0	0
c	a	0.357	0	0.482	0.161

DSP Phyto (lag = 1)	DSP Phyto(lag = 0)	P(ASP Phyto = a)	P(ASP Phyto = b)	P(ASP Phyto = c)	P(ASP Phyto = d)
d	a	1	0	0	0
a	d	1	0	0	0
d	d	1	0	0	0
a	a	0.987	0.003	0.007	0.003

DSP Phyto (lag = 1)	Chlorophyll(lag = 0)	P(DSP Phyto = a)	P(DSP Phyto = d)
d	a	0.500	0.500
a	c	1	0
a	b	0.993	0.007
a	a	1	0
d	d	0.500	0.500
d	c	0.500	0.500
d	b	0.333	0.667
a	d	1	0

DSP Phyto (lag = 1)	DSP Phyto(lag = 0)	P(PSP Phyto = a)	P(PSP Phyto = c)	P(PSP Phyto = d)
d	a	1	0	0
a	d	1	0	0
d	d	1	0	0
a	a	0.993	0.003	0.003

F

MAESTRO generated conditional probability tables for RIAV 2 zone

PSP Phyto (lag = 1)	PSP Phyto(lag = 0)	P(Lipophilic Toxins = a)	P(Lipophilic Toxins = c)	P(Lipophilic Toxins = b)
a	a	0.848	0.071	0.081
a	c	1	0	0
c	c	0.333	0.333	0.333
c	a	1	0	0

PSP Phyto (lag = 1)	PSP Phyto(lag = 0)	P(Amnesic Toxins = a)	P(Amnesic Toxins = b)	P(Amnesic Toxins = c)
a	c	1	0	0
a	a	0.985	0.005	0.010
c	c	0.333	0.333	0.333
c	a	1	0	0

PSP Phyto (lag = 1)	PSP Phyto(lag = 0)	P(Paralytic Toxins = a)	P(Paralytic Toxins = c)	P(Paralytic Toxins = b)
c	a	0	1	0
a	a	0.980	0.010	0.010
a	c	1	0	0
c	c	0.333	0.333	0.333

Chlorophyll (lag = 1)	PSP Phyto (lag = 0)	P(Chlorophyll = a)	P(Chlorophyll = b)	P(Chlorophyll = c)
a	c	1	0	0
a	a	0.795	0.189	0.016
c	c	0.333	0.333	0.333
b	a	0.391	0.578	0.031
c	a	0.125	0.375	0.500
b	c	0.333	0.333	0.333

SST (lag = 1)	P(SST = b)	P(SST = a)	P(SST = c)
b	0.659	0.143	0.198
a	0.150	0.850	0
c	0.621	0	0.379

ASP Phyto (lag = 1)	PSP Phyto (lag = 0)	P(ASP Phyto= a)	P(ASP Phyto = c)	P(ASP Phyto = b)
c	c	0.333	0.333	0.333
c	a	0	0	1
b	a	0.333	0	0.667
a	c	1	0	0
a	a	0.995	0.005	0
b	c	0.333	0.333	0.333

DSP Phyto (lag = 1)	PSP Phyto (lag = 0)	P(DSP Phyto= a)	P(DSP Phyto = c)	P(DSP Phyto = b)
c	a	0.111	0.889	0
a	a	0.989	0.005	0.005
b	c	0.333	0.333	0.333
b	a	1	0	0
a	c	1	0	0
c	c	0.333	0.333	0.333

PSP Phyto (lag = 1)	SST(lag = 0)	P(PSP Phyto = a)	P(PSP Phyto = c)
a	c	1	0
a	b	0.989	0.011
a	a	1	0
c	a	1	0
c	b	0.500	0.500
c	c	0.500	0.500

G

MAESTRO generated conditional probability tables for RIAV 3 zone

Amnesic Toxins (lag = 1)	Amnesic Toxins (lag = 0)	P(Lipophilic Toxins = a)	P(Lipophilic Toxins = b)	P(Lipophilic Toxins = c)
c	a	1	0	0
a	a	0.889	0.051	0.061
c	c	0.333	0.333	0.333
a	c	1	0	0

Amnesic Toxins (lag = 1)	ASP Phyto(lag = 0)	P(Amnesic Toxins = a)	P(Amnesic Toxins = c)
a	c	1	0
a	a	0.990	0.010
a	b	1	0
c	a	1	0
c	c	0.500	0.500
c	b	0.500	0.500

Amnesic Toxins (lag = 1)	Amnesic Toxins (lag = 0)	P(Paralytic Toxins = a)	P(Paralytic Toxins = c)	P(Paralytic Toxins = b)
c	c	0.333	0.333	0.333
a	a	0.975	0.010	0.015
a	c	1	0	0
c	a	1	0	0

Chlorophyll (lag = 1)	SST (lag = 0)	P(Chlorophyll = a)	P(Chlorophyll = b)	P(Chlorophyll = c)
c	a	0	1	0
b	a	0.650	0.350	0
b	b	0.342	0.605	0.053
b	c	0.300	0.400	0.300
c	b	0	0.400	0.600
c	c	0.250	0.250	0.500
a	a	0.917	0.083	0
a	b	0.571	0.429	0
a	c	0.400	0.600	0

SST (lag = 1)	Amnesic Toxins (lag = 0)	P(SST = a)	P(SST = b)	P(SST = c)
b	a	0.212	0.600	0.188
c	c	0	1	0
a	c	0	1	0
b	c	0.333	0.333	0.333
a	a	0.804	0.185	0.011
c	a	0.043	0.652	0.304

ASP Phyto (lag = 1)	P(ASP Phyto = a)	P(ASP Phyto = b)	P(ASP Phyto = c)
a	0.990	0.010	0
b	0.500	0	0.500
c	0.333	0	0.667

Amnesic Toxins (lag = 1)	Amnesic Toxins (lag = 0)	P(DSP Phyto= a)	P(DSP Phyto = b)	P(DSP Phyto = c)
c	a	0.500	0.500	0
a	a	0.985	0	0.015
c	c	0.333	0.333	0.333
a	c	1	0	0

Amnesic Toxins (lag = 1)	Amnesic Toxins (lag = 0)	P(PSP Phyto= a)	P(PSP Phyto = c)	P(PSP Phyto = b)
a	c	1	0	0
a	a	0.980	0.005	0.015
c	a	1	0	0
c	c	0.333	0.333	0.333

H

MAESTRO generated conditional probability tables for RIAV2 and RIAV3 timeseries (combined)

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Lipophilic Toxins R2 = a)	P(Lipophilic Toxins R2 = b)	P(Lipophilic Toxins R2 = c)	P(Lipophilic Toxins R2 = d)
d	a	1	0	0	0
a	d	1	0	0	0
a	a	0.806	0.056	0.102	0.036
d	d	0.250	0.250	0.250	0.250

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Amnesic Toxins R2 = a)	P(Amnesic Toxins R2 = b)	P(Amnesic Toxins R2 = d)
a	d	1	0	0
d	a	1	0	0
a	a	0.980	0.010	0.010
d	d	0.333	0.333	0.333

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Paralytic Toxins R2 = a)	P(Paralytic Toxins R2 = c)	P(Paralytic Toxins R2 = b)	P(Paralytic Toxins R2 = d)
d	d	0.250	0.250	0.250	0.250
d	a	0	1	0	0
a	a	0.975	0	0.015	0.010
a	d	1	0	0	0

DSP PhytoR2 (lag = 1)	ChlR3 (lag = 0)	P(Chlorophyll R2 = a)	P(Chlorophyll R2 = b)	P(Chlorophyll R2 = c)	P(ChlorophyllR2 = d)
a	d	0	0	0.333	0.667
d	a	0.250	0.250	0.250	0.250
a	c	0.174	0.391	0.304	0.130
d	b	0	1	0	0
d	d	0.250	0.250	0.250	0.250
a	a	0.778	0.222	0	0
a	b	0.264	0.604	0.132	0
d	c	0.250	0.250	0.250	0.250

PSP PhytoR2 (lag = 1)	SST R3 (lag = 0)	P(SST R2 = b)	P(SST R2 = a)	P(SST R2 = c)	P(SST R2 = d)
a	b	0.769	0.011	0.209	0.011
d	a	1	0	0	0
d	b	0.250	0.250	0.250	0.250
a	d	0.048	0	0.286	0.667
d	d	0.250	0.250	0.250	0.250
a	a	0.132	0.868	0	0
d	c	0.250	0.250	0.250	0.250
a	c	0.118	0	0.765	0.118

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(ASP PhytoR2 = a)	P(ASP PhytoR2 = d)	P(ASP PhytoR2 = c)	P(ASP PhytoR2 = b)
a	a	0.980	0.005	0.010	0.005
d	a	1	0	0	0
a	d	1	0	0	0
d	d	0.250	0.250	0.250	0.250

DSP PhytoR2 (lag = 1)	SST R3 (lag = 0)	P(DSP PhytoR2 = a)	P(DSP PhytoR2 = d)
d	a	0.500	0.500
d	b	1	0
d	d	0.500	0.500
d	c	1	0
a	b	0.989	0.011
a	a	1	0
a	d	1	0
a	c	0.970	0.030

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(PSP PhytoR2 = a)	P(PSP PhytoR2 = d
a	a	0.995	0.005
d	a	1	0
a	d	1	0
d	d	0.500	0.500

DSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Lipophilic Toxins R3 = a)	P(Lipophilic Toxins R3 = b)	P(Lipophilic Toxins R3 = c)	P(Lipophilic Toxins R3 = d)
d	a	1	0	0	0
a	d	1	0	0	0
d	d	0.250	0.250	0.250	0.250
a	a	0.832	0.082	0.041	0.046

DSP PhytoR2 (lag = 1)	Amnesic Toxins R2 (lag = 0)	P(Amnesic Toxins R3 = a)	P(Amnesic Toxins R3 = b)	P(Amnesic Toxins R3 = d)
a	a	1	0	0
d	d	0.333	0.333	0.333
d	a	1	0	0
d	b	0.333	0.333	0.333
a	d	0	0	1
a	b	0.500	0.500	0

PSP PhytoR2 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Paralytic Toxins R3 = a)	P(Paralytic Toxins R3 = d)	P(Paralytic Toxins R3 = b)	P(Paralytic Toxins R3 = c)
a	d	1	0	0	0
a	a	0.975	0	0.010	0.015
d	a	0	1	0	0
d	d	0.250	0.250	0.250	0.250

Chlorophyll R3 (lag = 1)	DSP Phyto R2 (lag = 0)	P(Chlorophyll R3 = a)	P(Chlorophyll R3 = b)	P(Chlorophyll R3 = c)	P(Chlorophyll R3 = d)
b	d	0	1	0	0
d	d	0.250	0.250	0.250	0.250
c	a	0.087	0.304	0.522	0.087
a	d	0	1	0	0
c	d	0.250	0.250	0.250	0.250
b	a	0.207	0.717	0.065	0.011
a	a	0.750	0.225	0.025	0
d	a	0	0	1	0

SST R3 (lag = 1)	P(SST R3 = b)	P(SST R3 = a)	P(SST R3 = d)	P(SST R3 = c)
d	0.238	0	0.476	0.286
c	0.382	0	0.206	0.412
a	0.151	0.849	0	0
b	0.707	0.098	0.043	0.152

ASP PhytoR3 (lag = 1)	DSP Phyto R2 (lag = 0)	P(ASP PhytoR3 = a)	P(ASP PhytoR3 = b)	P(ASP PhytoR3 = d)
a	a	0.990	0.010	0
d	d	0.333	0.333	0.333
d	a	0.250	0	0.750
a	d	1	0	0
b	d	0.333	0.333	0.333
b	a	0.333	0.333	0.333

DSP PhytoR2 (lag = 1)	DSP PhytoR2 (lag = 0)	P(DSP PhytoR3 = a)	P(DSP PhytoR3 = d)
d	d	0.5	0.5
a	d	1	0
d	a	1	0
a	a	0.5	0.5

DSP PhytoR2 (lag = 1)	PSP PhytoR2 (lag = 0)	P(PSP PhytoR3 = a)	P(PSP PhytoR3 = b)	P(PSPPhytoR3 = d)
d	d	0.333	0.333	0.333
a	d	0	1	0
a	a	0.995	0	0.005
d	a	1	0	0