Multivariate Time-Series Modeling of Shellfish Contamination with Dynamic Bayesian Networks

Michael S. Madeira
michael.madeira@tecnico.ulisboa.pt

Susana Vinga
susanavinga@tecnico.ulisboa.pt

Alexandra M. Carvalho
alexandra.carvalho@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2022

Abstract

Harmful algal blooms (HAB) create a natural contamination process of shellfish on production areas resulting from the accumulation of biotoxins produced by the HABs presence in high concentrations. Not only the public health problems have been the reason to develop predictive tools for this toxic events, but also the economic losses that currently affect many producers because of the variability of contamination on different shellfish species and harvesting areas. The goal was to find preliminary relations between the time series of the biotoxins concentration in shellfish and phytoplankton cell counts on water samples, between pairs of areas on a case study area in the south coast of Portugal, using correlation methods. The areas used for the analysis, that were on a potential geographic location to study the impact between each one, showed to have more informative time series in terms of its frequency of high toxic levels and the amount of records. A graphical model, named dynamic Bayesian network (DBN), was also built to extract the inter-timeslice dependencies, i.e. relation between the toxins and phytoplankton concentration at an area on a week with the concentration in the next week on its neighbor. The cross-correlation analysis showed that there is a higher positive relation from West to East in terms of phytoplankton dispersion, which is what is expected, but there was not clear evidence by analyzing the conditional probabilities, if there is an area that consistently explains what happens in the next week on its neighbors.

Keywords: Shellfish Contamination, Biotoxin, Harmful Algal Blooms, Time Series, Correlation, Dynamic Bayesian Networks

1. Introduction

Shellfish contamination generate negative impacts such on public health and food production sector economy, requiring monitoring programmes to regulate the opening and closures of the harvesting activities, to mitigate the health issues inflicted on humans, and the economic losses on shellfish farmers, aquaculture production, harvesters, and local businesses. There are different type of poisoning, and the five most commonly recognized are ciguatera fish poisoning (CFP), paralytic shellfish poisoning (PSP), neurotoxic shellfish poisoning (NSP), amnesic shellfish poisoning (ASP) and diarrhetic shellfish poisoning (DSP) [15]. According to the Portuguese monitoring program since 1985, the most reported HABs species are Pseudo-nitzschia spp., Dinophysis spp. and Gymnodinium spp. among others, that respectively produce biotoxins associated with ASP, DSP and PSP [8, 9].

An HAB and Shellfish harvesting warning bulletin was developed by IPMA to show current condition and one-week forecast of which production areas are open or close according to the biological regulatory values, the empirical know-how, the modelling of how sea surface temperature and tide directions influence algal proliferation and its position estimation through extracted variables like clorophyll-a [8]. In Ireland and Scotland were also implemented monitoring programmes with forecasting bulletins, that generate short-term predictions on the probability of occurrence of a toxic event, which can alert many aquaculture production areas

and other coastal natural harvesters [11]. This HAB warning solutions already brings deliberative abilities to manage opening and closures of harvesting areas, but considering the different abiotic and biotic factors that can be associated with HAB events, predictions by empirical rule-based models can be limited since it is normally underfitted to the specific areas and environmental conditions.

To complement the forecasting and to give more statistical assurance of how the forecasting results were the ones retrieved, a dependency modeling of the contamination process was developed as the main goal for this project. There are some works that already have tried to use multivariate models to forecast the toxin contamination event, some are able to have good results for a one week prediction [7, 11]. Another approach to this thesis was adopted under the MATISSE scope, a research project "MATISSE: A Machine Learning-Based Forecasting System for Shellfish Safety" (DSAIPA/DS/0026/2019), funded by the Foundation for Science and Technology, to focus more on modeling the target variables, which led me to get a model that could bring explainability to the event, by obtaining the dependencies between them and be an information and validation resource for multivariate forecasting models of what could be the specific and more relevant variables to help anticipate the contamination problem on Portugal coastal areas. Other goal was to model how the occurrence of DSP contamination at a shellfish production area could lead to the increase in toxins level or the phytoplankton concentration on other neighbor production areas. But what it was an additional and interesting analysis at first, ended as the center topic of this thesis.

The data available to give as input to the model, were collected and provided by IPMA (Portuguese Institute of Sea and Atmosphere), like biotoxins, phytoplankton and meteorological time series, by MARETEC, a research center with its main activity focusing on water environment data extraction and modeling, sea water hydrodynamic and nutrients time series, plus the sea surface temperature and clorophyll-a time series extracted from satellite imagery from the Copernicus program [1]. The target variables are collected from the the main inshore and offshore shellfish production and natural banks areas that are monitored weekly by IPMA, divided into 12 coastal areas, 7 estuaries, 3 lagoons and 16 rias, each one with different type of shellfish.

To model the dependency between the target variables (biotoxins and phytoplankton), a probabilistic graphical model named Dynamic Bayesian Network (DBN), will be used.

2. Background

2.1. Time Series Analysis

In time series forecasting it is also important to know the dependence between values of the series, by estimating autocorrelations, and to do that with precision, the series structure should be regular and not changing at every time step. However some models can model this components and get more precise predictions.

Auto correlation is commonly used to choose the important lag features on a variable's series, that on an auto regressive model could make a good prediction of the current value y_t. Lags are the sequence of values k time steps before the current time $t \in T$. Intuitively the autocorrelation, given the Eq. 1, is nothing less than the correlation or the degree of similarity between the current series and other lagged, usually being the Pearson coefficient.

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=1}^{T} (y_t - \overline{y})^2}. \quad (1)$$

Partial correlation also summarizes dependence on past observations, but only takes into account the direct effect or correlation of the chosen lag that could help on the auto regressive model to predict the series in current time step. It is calculated using the following Eq. 2:

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=k+1}^{T} (y_t - \overline{y})^2 \sum_{t=k+1}^{T} (y_{t-k} - \overline{y})^2}. \quad (2)$$

In this research context there are multiple variables observed over time, to be considered to model the biotoxins concentration and HAB events, and since each one may depend not only on its past values but also on other variables, the dependence between each pair of time series is important, and to calculate that, the cross-correlation is measured by using the same method of lagging time series, but now to calculate the correlation coefficient between two different time series, one stays fixed in time, and the second one is shifted backward or forward in time, as it is illustrated in the table below.

Table 1: Two Different Time Series Lagged

t	A	B	A	B (lag -1)	A	B (lag +1)
0	24	49	24	29	-	-
1	29	29	29	47	29	49
2	32	47	32	31	32	29
3	36	31	36	30	36	47
4	25	30	25	23	25	31
5	16	23	-	-	16	30

The correlation must be measured after the decomposition of deterministic time series components. Because trend make it seem like lags are very correlated. Detrended partial cross-correlation analysis (DPCCA) is another method to calculate cross-correlation, that also applies the partial correlation, removing the influence of the other lags, but it detrends both the time series when calculating their correlation. The correlation coefficient is calculated for multiple sub-series through a sliding window with a parameterized length s, which is then normalized by the coefficient of each time series. This correlation can characterize time dependent relation on different time scales [13].

2.2. Dynamic Bayesian Networks

DBNs are an extension of Bayesian Networks (BN), that can model both discrete and continuous variables, whose values change over time, and the network is able to propagate the probability distribution of the factors in the graph through time, that now turn probabilities into trajectories (an assignment of a value to each variable X_i at each time t).

To represent these multiple variables trajectories on a problem where the measurements are not in real time, and there is some time granularity in common, the timeline is discretized into time slices. As for the variables available in this problem, each time slice will correspond to 1 week. By considering the same conditional independence assumptions, but now within each time slice (intra-slice) and between each time slice (inter-slice) variables distributions, the probability of all random variables X along time T is:

$$P\left(X^{(0:T)}\right) = P\left(X^{(0)}\right) \prod_{t=0}^{T-1} P\left(X^{(t+1)} \middle| X^{(0:t)}\right) \tag{3}$$

the $P(X^{(0)})$ is the joint probability of the random variables at time step t=0, the initial distribution represented on a prior BN B_0, and the following conditional probabilities by a set of transition BNs $B_{\rightarrow}$. It defines a dependency of the joint distribution of X at a time step $t+1$ on the joint probability distribution of all variables' trajectories, within $[0 : t]$, that is infeasible to calculate since it is exponentially expensive to propagate such complex structures through time.

So usually a first simplifying Markov assumption is applied to only get conditional dependency on m timesteps before t. Simplifying eq:3:

$$P\left(X^{(0:T)}\right) = P\left(X^{(0)}\right) \prod_{t=0}^{T-1} P\left(X^{(t+1)} \middle| X^{(t-m:t)}\right) \tag{4}$$

A second simplification can be made depending on the problem in hands, i.e. having stationary transition networks between time slices or not, i.e. assuming time invariance if stationary implying the same transition model $P\left(X^{(t+1)} \middle| X^{(t)}\right)$ for all t.

DBNs are complex models with different possible BNs configurations, and make assumptions that may require appropriate model design, such as Markov assumption and Time Invariance. Therefore learning the structure and parameters of a DBN is also complex, since there are now intra-slice and inter slice dependencies to map. If the DBN follows the second assumption of being stationary, both structure and parameters are constant throughout time slices, and the intra-slice dependencies are always the same, and only the inter-slices dependencies must be learned. Still, if we want an optimal problem representation, both dependencies must be learned.

A fixed template transition model, based on the stationary property, and given the initial distribution of all variables, let us unroll the network over sequences of any length, i.e. it allow us to make inferences in the long-term notice.

The structure learning algorithm starts by initializing a complete directed graph, with bidirectional dependencies between all variables at timestep $t+1$, i.e. the intra-slice connectivity. Then rather than using mutual information to calculate the score for each edge, as in [2] where the edges are undirected and the score from a node $X_i[t+1]$ to $X_j[t+1]$ is equal in both directions, this adapted tree-augmented DBN learning algorithm uses a score that allows to learn intra and inter time-slice relations between variables to guarantee an optimal network, that expresses the gain in the total network score by including an intra-slice parent in $X[t+1]$ or only with parents from preceding time-slices $\mathbf{X}[t+1]$. But this does not restrict the complexity yet on the number of parents from the previous time-slices $\mathbf{X}[t]$, which could also be connected, so the maximum number of parents p from preceding time-slices must be set as well. The weight for each edge e_{ij} is then assigned as

$$e_{ij} = s_{ij} - s_i, \tag{5}$$

with the s_i being calculated as

$$s_i = \max_{X_p[t] \epsilon X[t]} \phi i(X_p[t], D_t^{t+1}), \tag{6}$$

where ϕi is the scoring function, and D_t^{t+1} is the fully observed data concerning the time transition $t \longrightarrow t+1$. So s_{ij} is calculated as

$$s_i = \max_{X_p[t] \epsilon X[t]} \phi i(X_p[t] \cup X_j[t+1], D_t^{t+1}). \tag{7}$$

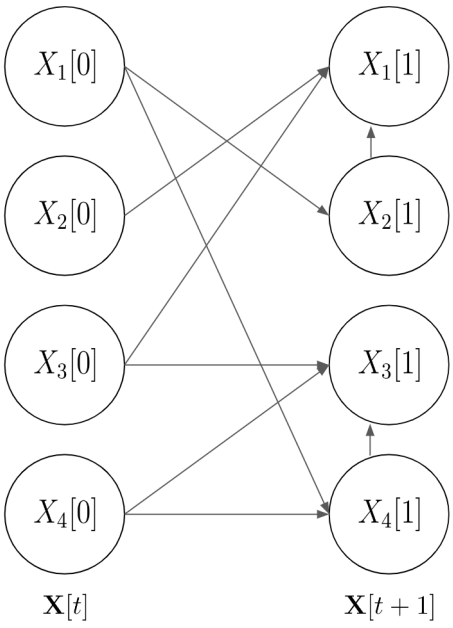


Figure 1: Learned DBN.

The scoring function used can be log-likelihood (LL) or the minimum description length (MDL), where the first one is usually prone to overfitting since its value increases proportionally to the number of parents added, since the entropy will never increase, so this score $LL(G|D) = -N \times ENT_D(X_1[1]|X_2[0], X_3[0], X_2[1]) + ENT_D(X_2[1]|X_1[0]) + ENT_D(X_3[1]|X_3[0], X_4[0], X_4[1]) + ENT_D(X_4[1]|X_1[0], X_4[0]) + ENT_D(X_1[0]) + ENT_D(X_2[0]) + ENT_D(X_3[0]) + ENT_D(X_4[0])$, will never decrease. And the second one uses a penalty factor proportionally to the number of parents added to the tree, $MDL(G|D) = LL(G|D) - \frac{1}{2}log(||D||).||G||$, G being the graph structure and D the dataset. The maximum spanning tree algorithm is then used to obtain the set of directed edges where the sum of all the scores is maximum, also removing whatever cyclical relations may exist, which defines the final directed acyclical graph, i.e. the stationary dynamic Bayesian network, such as in fig. 1.

The conditional entropy, assuming the set of discrete values n for the node $X_i[t+1]$ and the set q of the values combinations from the preceding timeslice parents $X_p[t]$ is given by

$$\mathrm{ENT}_D(X_i[t+1]|X_p[t]) = \sum_{i=1}^{n}\sum_{j=1}^{q} P(n_i, q_j) log \frac{P(n_i,q_j)}{P(n_i)P(q_j)}. \quad (8)$$

The DBN parameters, which are the conditional probabilities of each node, are calculated in a straightforward frequency calculation of events with each particular combination of values for each sub-DAG.

2.3. State of the Art

There has been studies focused on DBN framework to model different type of problems like for water eutrophication factors causality and inference [12, 3], and others for emerging HABs risk or biotoxins concentration like [4]. A non fully observed BN framework with an Hidden Markov Model structure, with water eutrophication level as the hidden variable, where each state affects the observed observation patterns composed by clorophyll-a concentration and a set of principal components resulting from a feature selection of abiotic variables like temperature, wind and others, has been used to forecast following values of a biotoxin [4]. It could be relevant as well, to map the causality and interrelations between contiguous areas, by applying an hierarquical BN representation [4], of coastal areas with multiple sampling points like estuaries or lagoons, and even between some of the bigger areas.

3. Implementation

The type of project requires an whole process of getting to know the data beforehand, understand each data source to then integrate it on a standard format to then apply multivariate analysis and modeling. To perform the analysis proposed every time series, collected from or provided by IPMA, MARETEC and Copernicus satellite programme, went through a process of data cleaning and integration.

3.1. Data Cleaning and Integration

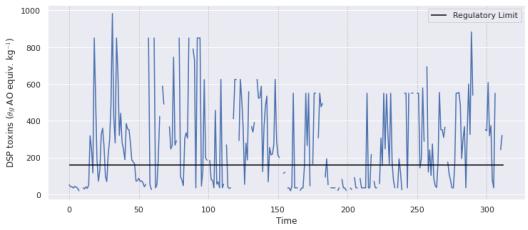
IPMA's biological time series, such as the biotoxins collected in-situ from shellfish tissue samples and the phytoplankton cells count within water samples, needed a lot of records values cleaning and replacement, in terms of some specific indicative sampling results, like non-detectable (ND), non-quantifiable (NQ) or even not registered (NR), that were replaced by the toxin quantification detectable limit, like 36 µg/kg for DSP toxins, 1.8 µg/kg for ASP, and 71 µg/kg for PSP, and the "NR" values were replaced by "n/a". Also for some typos and mislabelled production areas, sampling stations and species names that needed to be standardized, in order to select the time series individually. The phytoplankton time series records for each production area, went through the same process of names typos correction and the not numerical interpretative values, like "LD", now representing samples with cell counts too low to be detected, were replaced by the a detectable limit value of 20 cel/L provided by IPMA. The same process for IPMA's meteorological time series were applied, only with different values to be replaced.

For the data provided by MARETEC and ex-

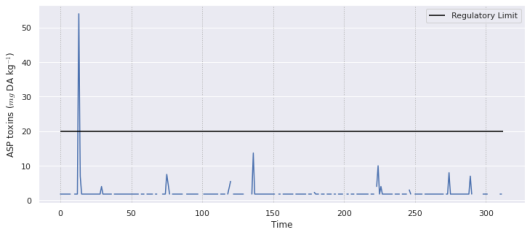
tracted from Copernicus satellite products, the main task was to integrate and merge with the other time series, according to the coarser granularity of the biological data. So given an expected weekly frequency, the IPMA biological data went also through a process of reindexing the records to a precise weekly granularity, where a new weekly frequency index is created for both time series, from 2015/01/01 to the 2020/12/31 (the data available period), where the weekly record value will now be the rolling mean of the 7 days around the new index date in the center, so it aggregates the time series values from 3 days before to 3 days after and apply the average value, because there clearly some periods where some records were taken on more sparse periods of time, and others on a more rigid scheduled regime. Consequently the meteorological data from IPMA, the oceanographic and water nutrients data from MARETEC, and the sea surface temperature and clorophyll-a from Copernicus remote imagery data, followed this record frequency, passing to a daily or even hourly frequency to a weekly one, which required an aggregation for the finest granularity, and a simpler merging approach of the closest day given the day of the week of the biological records, to give a more realistic information since this are not time series in real time.

3.2. Experiment Definition and Design

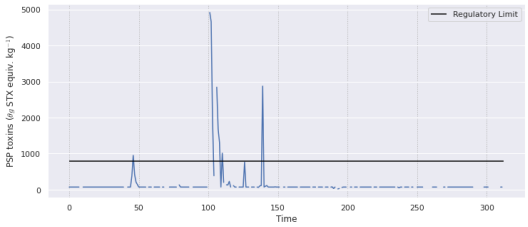
To reduce the analysis and modeling complexity of the number of production areas, sampling points and different species, a ranking method, was implemented to empirically select the most informative entities (composed as production area — sampling point — specie), in terms of the time series with higher records frequency with high toxicity, higher phytoplankton agglomeration and lower percentage of missing values. First of all, a filtering threshold was applied to select only the entities with time series that had at least 5% as minimum percentage of high toxicity events, which means that at least five times out of 100 toxin records the concentration would surpass the respective toxin limit (160 µg/kg for DSP, 20 µg/kg for ASP, 800 µg/kg for PSP), a percentage of phytoplankton concentration higher than 5%, which means getting more than 200 cel/L, and also less than 20% missing values, to diminish the amount of synthetic data added in the imputation process. However to get the analysis between each entity another threshold was implemented in order to get at least 104 records, i.e. 2 years, with the previous settled conditions. In the end, 20 entities were considered the relevant ones, from which 1 test scenario was chosen to explore the hypothetical predictive linkage of an area with its neighbors. Vale da Lama (LAG), Fortaleza (OLH2), Quatro Águas (TAV) and Monte Gordo (L9) from the south coast of Continental Portugal, did fill the preliminary requirements, but also add an extra aspect of geographical proximity, and maritime conditions, since there is evidence of the direction of the sea surface current at south coast of Continental Portugal being broadly originated from the Azores current extended to the Gulf of Cadiz, thus having a eastward direction, i.e. from West to East [5].



(a)



(b)



(c)

Figure 2: Toxins Time series of Ria de Aveiro, Piscicultura

After this first entities selection there was also a defined focus in terms of the specie and biotoxin analysis scope. Mussels were considered to be the ones that got more records over all entities, and as the specie is considered to be the most indicatory of a contamination event, regarding its capacity of accumulating biotoxin easily and faster comparing to the elimination process [9, 10], it. DSP biotoxins also revealed to be the more dominant one, generally, as it is represented in fig. 2 in Ria de Aveiro (RIAV1), as an example, but also supported by this study [6], as the toxin most common in the Portuguese maritime coast.

From this point, it was designed a case study to analyze the impact that each area, DSP toxins concentration on shellfish and phytoplankton on water, has in the following weeks on the concentrations of its neighbors, by calculating the cross-correlation

coefficient within the range of a lagged time-window settled between -8 and +8 lags, identify the higher peak and the direction of that lag, and find preliminary relations between the time series of each pair of areas, through distinct methods, known as Pearson correlation and DPCCA, already explained in section 2.

Since a higher correlation doesn't mean a certain dependency relation between both random variables, a model to explain the multivariate intertimeslice dependencies was required. The graphical model web tool used to built the DBNs to model this contamination dependency scenarios and compress the knowledge on a graph was MAESTRO, and it can be accessed in https://vascocandeias.github.io/maestro/.

3.3. Data Pre-Processing

To find model parameters and Bayesian network structure, and train it, some data requirements already described in section 2 must be fulfilled. Although the ability of Bayesian models to manage a good model fit to data even with missing records, this specific Bayesian architecture has a learning algorithm that requires full observability, which asks for a filling or replacement strategy of missing values. But at the same time it is common to reduce as much as possible the bias added to the dependency interpretation that may explain the contamination phenomenon by toxic algae between neighboring areas, on the imputation process of an unnecessary amount of synthetic data, so that's why it is important to study time series completeness of the relevant areas and its neighbors. The second requirement is limiting the continuous values to a number of pre-defined values, through the time series discretization.

First and foremost a best time period in common for each pair was selected, fulfilling the same requirements of an entity considered informative, to reduce the bias that will come from filling missing values with synthetic data, and also give richer time series to better model the dependencies, since the toxic contamination events are extremely rare.

The missing values from phytoplankton and mostly toxins may come from the fact that there are no need for sampling in-situ during the period that the area is closed, since there are some rules of thumb for how much weeks it should be closed. So two methods were used consecutively to fill the values with information aggregated from a rolling average from 3 weeks before to 3 weeks after the date of the missing record, and then the ones that did happen on a bigger period with no weekly sampling were filled on a more conservative way, i.e. by their default value of detectable limit.

Then the DSP toxin time series is divided in three levels, the first level between 0 and 10% below the DSP limit value ([0-144]), the second level around the limit, between 10% below the DSP limit value and 10% above the DSP limit value ([144-176]) and the third level between 176 and the maximum value, being represented by 0, 144 and 176 as the respective labels for each interval. And the DSP phytoplankton also was discretized into three levels, a small concentration of phytoplankton cells per liter between the phytoplankton detectable limit of 20 and the toxic threshold of 200, excluding 200, a medium concentration level between 200 and 2000, and a higher concentration from 2000 to the maximum value, represented by 20, 200 and 2000 as the respective interval labels.

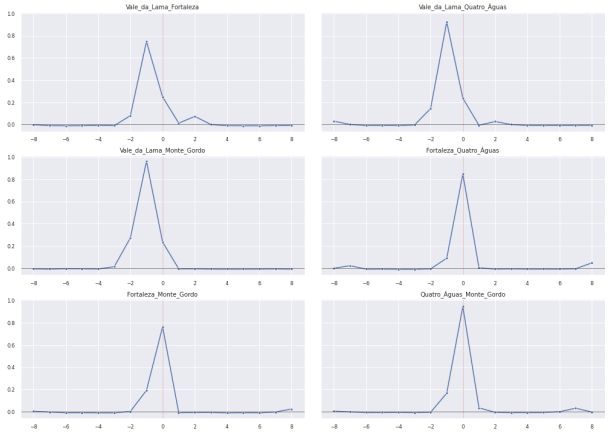
3.4. Model Training

As the data already had been pre-processed, the MAESTRO tool will only need to configure the learning or training process of the DBN. There is the option of learning a stationary or a non-stationary DBN network, but since the time series have more than 300 time-slices or recording events, the stationary network is more appropriate to model this empirical data distribution. To learn the stationary Bayesian structure, some parameters must be set, like the number of markov lags, which empirically and for simpler experiment purposes can be defined by 1 or 2 markov lags, 1 or 2 weeks respectively, in this case study 1 week between areas was assumed. The scoring function can vary between the log-likelihood (LL) or the minimum description length (MDL), where the LL score is preferred over the MDL, because the way MDL penalizes the structure learning process in terms of adding parents, it forces the network to have only one parent per node, and as it will be found on cross-correlation analysis that one variable may not be able to fully explain the variation of another one, and the LL scoring function limited by the maximum number of parents from preceding time-slice(s) may show what other variables could help to get more explainability. The maximum number of parents were defined as 2 for preceding time-slice parents, to help optimize the structure but mainly to focus on the more informative interdependencies. Finally the inter-slice relations between the same variable has been forbidden by passing a comma separated values file with this format example per variable (1,dsp toxin Vale da Lama,dsp toxin Vale da Lama,-1), allowing for other relations to stand out, since from some experiments this were the most returned, but assuming this is an empirical filter this must be optimized in future work.

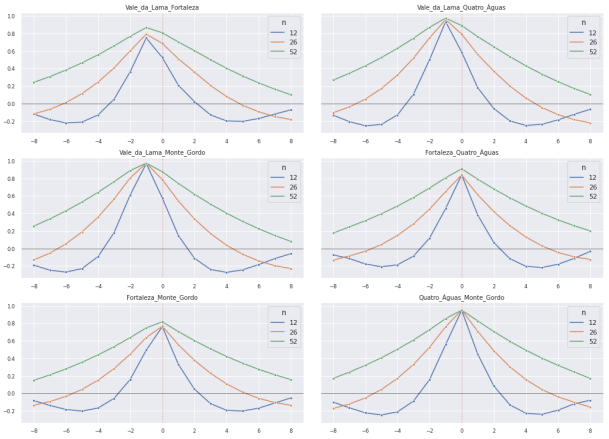
4. Results

4.1. Cross-Correlation Results

The correlation analysis showed that there are statistical evidence that for almost every pair of areas the correlation peaks are around the same timeslice, i.e. lag 0, or within the range of -2 and 2 weekly lags, which is important to notice that exist some relation between what happens a week before on an area and a week after on a contiguous or more distant area. Every time series pair phytoplankton-phytoplankton between LAG and the other areas are presented with a high positive coefficient on the lag -1, meaning that in general the phytoplankton at all the case study areas have a directly proportional growth relatively to the phytoplankton at Vale da Lama on the previous week, as it can be observed in fig. 3.



(a)



(b)

Figure 3: South Case Study Phyto-Phyto Pearson Correlations (a), and DPCCA Correlations (b).

By analysing the DSP toxins the correlation between DSP toxins at Quatro Águas and Fortaleza DSP toxins apart from having a lower Pearson correlation of around 0.4 on both lag -1 and +1, was the highest one among all pairs, but generally there were higher correlations on lag +1 which can be kind of counter intuitive, since the proliferation of HAB and its dispersion should be gradual, and one area could be impacted first, meaning that the toxin concentration on shellfish on both production areas gets high at the same time and there is no sequence of toxin effect.

4.2. Dynamic Bayesian Models Results

After obtaining the indicative relations from the cross-correlation analysis, the idea was to extract from the dynamic Bayesian networks what are the most probable conditions that helps the variable of interest to get higher, i.e. an increase on toxin or phytoplankton concentration, depending on the time series relation. To confirm the relations that the cross-correlation coefficients already pointed out to be high on 1 or 2 lagged time series, the minimum requirement is that at least that inter-timeslice dependency was learned by the DBN structure as one of the variables that contribute to the explanation of the variation of the variable in the next timeslice. If the correlation coefficient of this relation was high as it shows to be in the LAG-TAV time series cross-correlation, it means that the variable at lag -1 explains the other time series growth, i.e. the phytoplankton at TAV has a predictive linkage with the phytoplankton at LAG in the previous week, thus the expected dependency being like $dsp\_phyto\_Vale\_da\_Lama[0] \longrightarrow dsp\_phyto\_Quatro\_Águas[1]$, but not completely otherwise it would be a Pearson coefficient equal to 1, so another variable could help explain its growth over time, and it can be discovered in the Bayesian dependency model. If the relation had a lower coefficient, but indicative that would be dependent, it is expected to be noticed a higher number of dependencies on other inter-slice variables as well. Since none of the variables were capable of having a correlation equal to 1, the DBN learned a sub-tree for each variable in the final state of the static network, and these should have a maximum of 2 parents ideally from the inter-slices and 1 from the intra-slice as default, by using the log-likelihood as the scoring function, as it is observed on the LAG-TAV DBN in fig. 4.

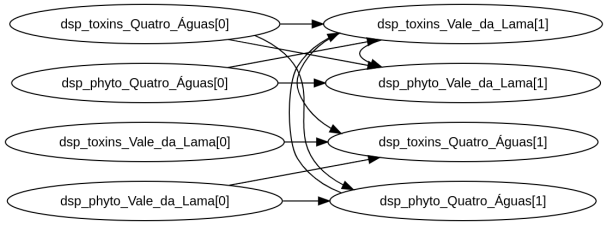


Figure 4: LAG-TAV DBN

P(dsp_toxins_Quatro_Águas[1] = 144)	P(dsp_toxins_Quatro_Águas[1] = 176)	dsp_toxins_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]
0.047	0.109	36	20	36
0	0.125	144	20	36
0.111	0.111	176	20	36
0	0	36	200	36
0	0	144	200	36
0.333	0.333	176	200	36
0.5	0	36	2000	36
0	0	144	2000	36
0	1	176	2000	36

(a)

P(dsp_toxins_Vale_da_Lama[1] = 144)	P(dsp_toxins_Vale_da_Lama[1] = 176)	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]
0.023	0.058	36	20	20
0	0	144	20	20
0.074	0.074	176	20	20
0.125	0	36	200	20
0	0	144	200	20
0.25	0.25	176	200	20
0	0	36	2000	20
0.333	0.333	144	2000	20
1	0	176	2000	20

(b)

P(dsp_phyto_Quatro_Águas[1] = 200)	P(dsp_phyto_Quatro_Águas[1] = 2000)	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Vale_da_Lama[0]
0.079	0.01	36	20
0.143	0	36	200
0.25	0.25	36	2000
0.3	0	144	20
0.333	0.333	144	200
0	1	144	2000

(c)

Figure 5: Two Sub-CPTs of DSP toxins at Quatro Águas (a) and at Vale da Lama (b), and the Sub-CPT of DSP phyto at Quatro Águas (c) in LAG-TAV DBN

To analyze the probabilities and extract insights that could confirm the indicative predictive linkage from the cross-correlation analysis, an interpretation model was designed, to confirm or not, that relation or dependency. First and foremost I try to lock the variable in the same timeslice on a conditional probability table (CPT) of a variable of interest like *dsp_phyto_Quatro_Águas*[1] and look for conditional probabilities of the higher toxins levels (144 and 176) or phytoplankton (200 and 2000) concentrations, higher than the probability of the low level concentration. On the sub-CPTs with high probabilities on the higher concentration states, it is also expected that the conditional probability of toxin or phytoplankton concentration on an area would increase proportionally as its parents states from the preceding time step gets higher as well, especially the one analyzed in cross-correlation, proving that the preliminary positive correlation with the variable from the neighboring area did indicate some prior degree of dependency, but for the other variables it may vary. So if the conditional probabilities were not as high as it was expected, it is important to show signs of proportional growth with its parents from the preceding timeslices, and preferentially by isolating the intra-slice parent, otherwise it is not an interesting dependency analysis for a predictive tool based on past information. Meaning that the objetive is to find a possible impact of each area toxin or phytoplankton concentration growth on the probability of occurring toxic events on their neighbors, may be shown on the increase in the probability of having a higher toxin or phytoplankton concentration when it was high on past weeks on its neighbors, and then conclude based on that conditional probability increase that a DSP toxin or phytoplankton measurement on a certain area can potentially act as an early warning alert with some probability.

To isolate the impact of the variable in the same time step as the target variable and focus on the probability growth and only look at the change in states of the other variables from the preceding timeslices, I need to nullify the other variable changing state, and to do that I can look to the CPT table as having sub-CPTs in it, and for each set of state combinations between the other preceding variables, the variable to be blocked always have the

same value, i.e. it does not change its state. If there are some higher probabilities on higher toxin or phytoplankton concentration when there is a lower concentration on the variable that was expected to have a higher positive correlation, could mean the contrary which means that the correlation should be negative, since when one increases the other decreases. But it can simply be the events that happen less frequently and don't contribute for a higher cross-correlation, so there is not a clear and enough evidence of a constant increase of toxins or phytoplankton in that direction, meaning that other areas and environmental factors may help on the proliferation and contamination as well.

In Vale da Lama (LAG) and Quatro Águas (TAV) DBN, the more expected inter-slice dependencies were from *dsp_phyto_Vale_da_Lama*[0] to *dsp_phyto_Quatro_Águas*[1], and from *dsp_toxins_Quatro_Águas*[0] to *dsp_toxins_Vale_da_Lama*[1], and the DBN structure confirms that dependencies. By analyzing the *dsp_toxins_Vale_da_Lama*[1] and the *dsp_toxins_Quatro_Águas*[1] node sub-CPTs of the CPTs, the sub-CPT in fig. 5 where the isolated variable is in the lowest state, shows that the probability of DSP toxin concentration at Quatro Águas (TAV) above 176 µg/kg increases upon the increase on both DSP toxins and DSP phytoplankton concentration at Vale da Lama (LAG), on the week before (1 lag). And on the opposite direction, the probability of DSP toxin concentration at Vale da Lama (LAG) between 144 µg/kg and 176 µg/kg, i.e the discretized state equal to 144, there is also an increase upon the increase in both DSP toxins and DSP phytoplankton concentration at Quatro Águas (TAV).

So it seems that the DBN for this areas pair suggests that there is no clear dominance in terms of impact on either direction (LAG $\longrightarrow$ TAV or (TAV $\longrightarrow$ LAG). The *dsp_phyto_Quatro_Águas[1]* already show a very high correlation with *dsp_phyto_Vale_da_Lama[0]*, but the dependence with *dsp_toxins_Quatro_Águas[0]* is obviously completing the explainability of the growth of phytoplankton at Quatro Águas, since on both states above the limit (200 and 2000), there is an increase in the conditional probability only by increasing the phytoplankton state at Vale da Lama, and only when the toxins at Quatro Águas state in the next week is of 2000, getting a probability equal to 1.

The same analysis was performed for all the pairs, and it was not possible to assume with high certainty the impact between the areas in the south coast, but the expected sea current direction in terms of phytoplankton dispersion was more or less confirmed.

5. Conclusions

This project comes as a complement for regression predictive models, as a feature selection tool for an inference model for some specific areas that may need specific attributes given as input to learn a regression model, a traditional or a deep learning one, and validate those models by giving an interpretative perspective. The Dynamic Bayesian networks built together with the cross-correlation analysis, not only proves the degree of dependency and correlation between the time series on different areas, but also it translates what areas impact on a toxic contamination occurrence more frequently and could indicate with 1 week notice in what areas could happen as well with a probability associated. In conclusion this results can be further explored and improved, to provide a more informative tool to solve the economic problem on all the chain of shellfish harvesting activities, by including the MARETEC phytoplankton from MOHID model or the clorophyll-a from Copernicus data can have potential to get more offshore information, and define different intermediate locations that could influence the more in-situ variables of IPMA such as phytoplankton and toxins.

Acknowledgements

I would like to thank my supervisors, Professor Alexandra Carvalho and Professor Susana Vinga, and my special adviser Doctor Marta Lopes for their weekly continuous support since the beginning, the feedback, encouragement and all the shared knowledge, especially through the last mile. Also I would like to thank for the contribution of my colleague Rafaela Cruz, André Patrício, and all the students that contributed in this project. Last but not least, to all my family that always supported me through my process as a student and person, and of course my friends and teammates without them it would all be more difficult.

Thank you all.

References

[1] https://www.copernicus.eu/en - European Union's Earth Observation Programme.

[2] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[3] A. Gudimov, E. O'Connor, M. Dittrich, H. Jarjanazi, M. E. Palmer, E. Stainsby, J. G. Winter, J. D. Young, and G. B. Arhonditsis. Continuous Bayesian network for studying the causal links between phosphorus loading and plankton patterns in lake simcoe, ontario, canada. *Environmental Sci-*

ence & Technology, 46(13):7283–7292, 2012. [https://doi.org/10.1021/es300983rCrossRef].

[4] P. Jiang, X. Liu, J. Zhang, and X. Yuan. A framework based on hidden Markov model with adaptive weighting for microcystin forecasting and early-warning. *Decision Support Systems*, 84:89–103, 2016. [https://doi.org/10.1016/j.dss.2016.02.003CrossRef].

[5] C. Martins, M. Sena-Martins, and A. Fiúza. Surface circulation in the eastern north atlantic, from drifters and altimetry. *J. Geophys. Res*, 107, 12 2002.

[6] R. Salas and D. Clarke. Review of dsp toxicity in ireland: Long-term trend impacts, biodiversity and toxin profiles from a monitoring perspective. *Toxins*, 11(2), 2019.

[7] W. Schmidt, H. L. Evers-King, C. J. A. Campos, D. B. Jones, P. I. Miller, K. Davidson, and J. D. Shutler. A generic approach for the development of short-term predictions of escherichia coli and biotoxins in shellfish. *Aquaculture Environment Interactions*, 10, 2018.

[8] A. Silva, L. Pinto, S. Rodrigues, H. de Pablo, M. Santos, T. Moita, and M. Mateus. A hab warning system for shellfish harvesting in portugal. *Harmful Algae*, 53:33–39, 2016. Applied Simulations and Integrated Modelling for the Understanding of Toxic and Harmful Algal Blooms (ASIMUTH).

[9] P. Vale, M. J. Botelho, S. M. Rodrigues, S. S. Gomes, and M. A. de M. Sampayo. Two decades of marine biotoxin monitoring in bivalves from portugal (1986–2006): A review of exposure assessment. *Harmful Algae*, 7(1):11–25, 2008. [https://doi.org/10.1016/j.hal.2015.11.017CrossRef].

[10] P. Vale and M. A. de M. Sampayo. Seasonality of diarrhetic shellfish poisoning at a coastal lagoon in portugal: rainfall patterns and folk wisdom. *Toxicon*, 41(2):187–197, 2003.

[11] X. Wang, Y. Bouzembrak, H. J. Marvin, D. Clarke, and F. Butler. Bayesian networks modeling of diarrhetic shellfish poisoning in mytilus edulis harvested in bantry bay, ireland. *Harmful Algae*, 112:102171, 2022.

[12] Z. Wu, Y. Liu, Z. Liang, S. Wu, and H. Guo. Internal cycling, not external loading, decides the nutrient limitation in eutrophic lake: A dynamic model with temporal Bayesian hierarchical inference. *Water Research*, 116:231–240, 06 2017. [https://doi.org/10.1016/j.watres.2017.03.039CrossRef].

[13] N. Yuan, Z. Fu, H. Zhang, L. Piao, E. Xoplaki, and J. Luterbacher. Detrended partial-cross-correlation analysis: a new method for analyzing correlations in complex system. *Scientific reports*, 5(1):1–7, 2015.