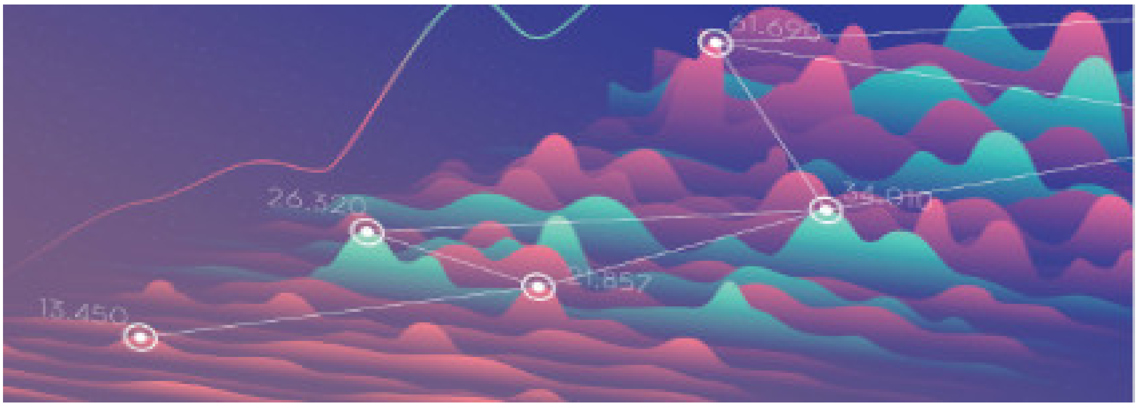
TÉCNICO
LISBOA



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

Michael Sousa Madeira

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Susana de Almeida Mendes Vinga Martins
Prof. Alexandra Sofia Martins de Carvalho

Examination Committee

Chairperson: Prof. Daniel Jorge Viegas Gonçalves
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Bruno Emanuel da Graça Martins

June 2022

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

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.

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 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

Resumo

As aglomerações de algas nocivas, criam um processo natural de contaminação de mariscos nas áreas de produção, resultado da acumulação das biotoxinas produzidas pelas altas concentrações de algas nocivas. Não só os problemas de saúde pública, vieram trazer a necessidade de se criarem ferramentas preditivas para estes eventos tóxicos, mas também os problemas e perdas económicos que atualmente afetam os demais produtores por causa da variabilidade da contaminação, nas diferentes espécies e nas diferentes áreas de apanha de marisco. O objetivo era encontrar relações preliminares entre as séries temporais das concentrações das toxinas e phytoplankton, entre os pares de áreas dentro da área de estudo do sul da costa Portuguesa, usando métodos de correlação. As áreas usadas para a análise, que estão numa localização geográfica potencial para serem estudadas, mostraram ser das mais informativas em termos de frequência de níveis de toxicidade altos e da quantidade de registos. Um modelo de grafos, chamado de rede Bayesiana dinâmica, foi também desenvolvido para extrair as dependências inter-temporais, isto é a relação entre concentração das toxinas e phytoplankton de uma área numa semana, e a concentração na semana seguinte de uma área vizinha. A análise da correlação cruzada mostrou que existe uma alta relação positiva de Oeste para Este em termos de dispersão de phytoplankton, o que é o esperado, mas não houve evidências claras ao analisar as probabilidades condicionadas, que haja uma área que de forma consistente possa explicar, o que acontece na semana seguinte nas áreas vizinhas.

Palavras Chave

Contaminação de Marisco, Biotoxinas, Aglomerações de Algas Nocivas, Séries Temporais, Correlação, Redes Bayesianas Dinâmicas

Contents

List of Figures

List of Tables

Acronyms

ASP Amnesic Shellfish Poisoning

CHL-A Clorophyll-a

CPT Conditional Probability Table

DBN Dynamic Bayesian Network

DPCCA Detrended Partial Cross Correlation Analysis

DSP Diarrhetic Shellfish Poisoning

HAB Harmful Algal Blooms

IPMA Portuguese Institute of Sea and Atmosphere

LL Log-Likelihood

MDL Minimum Description Length

MTS Multivariate Time Series

PSP Paralytic Shellfish Poisoning

SST Sea Surface Temperature

Chapter 1

Introduction

1.1 Problem Statement

Shellfish contamination is a great risk to public health, given shellfish is a human food source and its demand is rising [1], requiring more production and harvesting regulations and monitoring systems regarding consumer protection. Public authorities have implemented management plans with periodic measurements on contamination of shellfish by different biotoxins produced by harmful microalgae, bacteria, e.g. *Escherichia coli* (*E.coli*), and others, that within acceptable threshold values according to the species, will allow the opening of harvesting or its closure otherwise [2].

Harmful algal blooms (HAB) are a natural part of the seasonal cycle of photosynthetic organisms in marine ecosystems, resulting on fast proliferation of toxic phytoplankton. There has been many studies on how much the uprising consequences of global warming and eutrophication are directly and undirectly related with this HAB events [3–5]. For the time these events occur, shellfish may accumulate or eliminate biotoxins contamination, depending on each species metabolic capabilities, and still remain contaminated and unsafe for human consumption [6].

Water eutrophication, an increase in the rate of supply of organic matter to an ecosystem [7], results in excessive phytoplankton growth, low dissolved oxygen, degeneration of submerged macrophytes, and increased frequency algae blooms on coastal waters [8]. Normally it begins when there is an unbalanced and exceeded presence of nutrients like nitrogen and phosphorus, since phytoplankton feeds from them, and consequently helps to proliferate algal blooms [9]. And this nutrients' condition variation is related with many different factors, such as the use of sewage or waste water treatment technologies, on fertilizer application, on precipitation patterns, and on freshwater discharge [5].

Global Warming have shown evidence of temperature rising in the last decades. The effect of temperature on phytoplankton physiology and metabolic processes is well known. First, under light-saturated conditions, higher temperature increases specific phytoplankton productivity by acting on photosynthetic carbon assimilation. In addition, under non-limiting nutrient conditions, an increase in water temperature increases phytoplankton nutrient uptake and consequently its growth and cell multiplication. [10]. In Portugal the period in which high levels of toxins start to occur more frequently is around June and August, and generally around Europe, in the autumn and winter season there is a tendency of registering higher toxin accumulation in shellfish due to lower elimination, i.e. shellfish turn out to get a slower metabolism [11, 12].

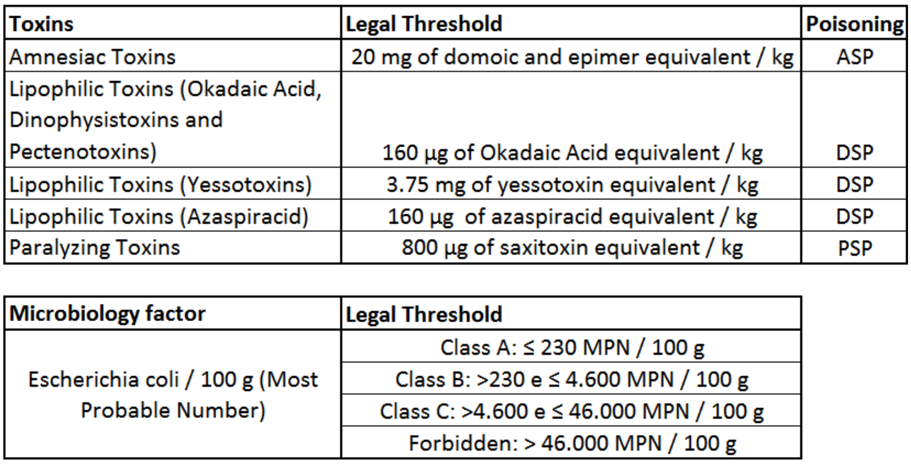
Not all microalgae species are harmful, and its majority contribute to shellfish own nourishment and in some cases there is a beneficial presence because it can significantly reduce faecal contamination [13]. *E. coli* is a bacterium that when present indicates a faecal microbiological contamination, since it is most abundant on human intestinal microflora, and overall is widely distributed in the intestine of warmblooded animals. Depending on the strain it can cause gastrointestinal, urinary or nervous diseases, mainly when the fish or shellfish are not properly cooked [14].

Thus it seems that a balance in the presence of many abiotic (non-living components in an ecosystem like physical conditions) and biotic (living beings present in an ecosystem) factors are essential to bring some ecological equilibrium. Nonetheless, phytoplankton species that are capable of producing marine biotoxins, are still a serious public health hazard and the main challenge is identifying harmful phytoplankton, originated in different climate conditions and from different geographic origins, in order to anticipate a biotoxin contamination on such different coastal locations, like estuarines, lagoons and even more offshore areas with different oceanographic characteristics, with the year to year variability of seasonal toxic peaks.

1.2 Motivation

Contamination by biotoxins can generate innumerous health issues through food intoxication with 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 toxins associated with ASP, DSP and PSP [12, 16].

There are already reactive measures to control the shellfish that goes into the public market or not, and the harvesting closure moment, based on legal thresholds for biotoxins and *E. coli* concentrations established by European Regulations [17, 18], adapted and monitored by the Portuguese Institute for the Sea and Atmosphere (IPMA), as showned in Fig. 1.1. But apart from possible errors on measurements and classification, it would be highly desirable to deliberately make the decision of not harvesting with anticipation, in order to prevent losses on shellfish farmers from aquaculture production, harvesters and consequently the economy in the affected regions. And even a higher entity such as maritime police, should have tools to apply jurisdiction measures on fishing practices on a early basis warnings.



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

Microbiology factor	Legal Threshold
Escherichia coli / 100 g (Most Probable Number)	Class A: ≤ 230 MPN / 100 g
	Class B: >230 e ≤ 4.600 MPN / 100 g
	Class C: >4.600 e ≤ 46.000 MPN / 100 g
	Forbidden: > 46.000 MPN / 100 g

Figure 1.1: Legal thresholds of main biotic factors on shellfish contamination. Adapted and translated from IPMA; Source: http://www.ipma.pt/pt/shellfish/docs/index.jsp

In Portugal, the shellfish harvesting business just in Ria de Aveiro had a profit around €6 million, in 2021, and since its an area with set of lagoons there is natural accumulation of phytoplankton. The shellfish production areas have regional variability in terms of shellfish species, but some species dominate the production markets, such as mussels and oysters. Only in 2020 the Irish shellfish production industry has an estimated profit total of €51 million from around 24,000 tons [19], which show the big dimension of this activity. The main inshore and offshore shellfish production and natural banks areas that are monitored weekly by IPMA are divided into 12 coastal areas, 7 estuaries, 3 lagoons and 16 rias, each one with different type of shellfish, represented in fig. 1.2. Multiple sampling stations and environmental stations, extract data in-situ, such as toxins and faecal indicator concentration on shellfish and phytoplankton cells density, temperature, wind and rainfall levels. The information available is significantly enriched, including remote sensing data, such as sea surface temperature and clorophyll-a concentration from satellite images taken of Copernicus program [20], and hydrodynamic and water nutrients generated by the operational oceeonographical model MOHID that simulate the transport of HABs alongshore [21]. 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 [16].

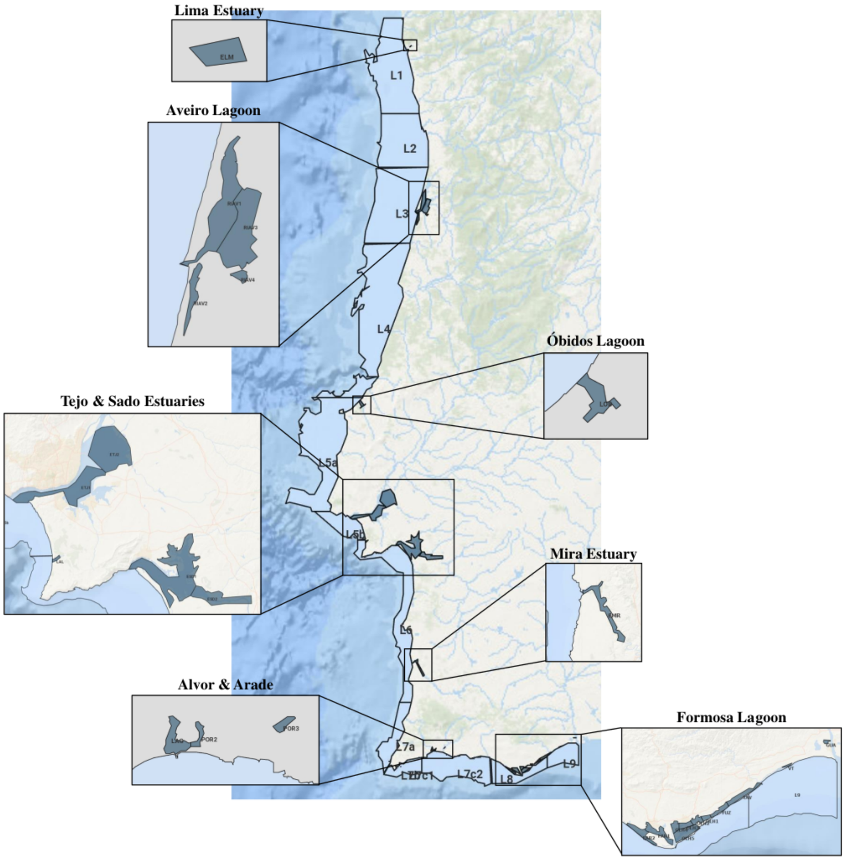


Figure 1.2: Shellfish production areas across Portuguese coast as defined by IPMA (adapted for IPMA`https://www.ipma.pt/pt/bivalves/index.jsp`)

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 [22]. 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. Machine learning (ML) models can handle complex correlation between such time-dependent variables, generalize the model as much as possible by continuously retraining, to predict shellfish contamination with robust learning from historical data or even model the phenomenon causality.

1.3 Main Goals

This thesis was developed under the scope of research project "MATISSE: A Machine Learning-Based Forecasting System for Shellfish Safety" (DSAIPA/DS/0026/2019), funded by the Foundation for Science and Technology, which aims to mitigate economic losses on this sector, by anticipating the closure and opening periods.

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 that is ultimately what want to forecast, some are able to have good results for a one week prediction [22,23]. Another approach to this thesis was adopted under the MATISSE scope, to focus more on modeling the target variables, which led me to get a model that could bring explainability to the event, by modeling the dependencies between them and inform multivariate forecasting models of what could be the specific target for 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.

As any other machine learning project, it can be divided into multiple different tasks. The first task and maybe the most important one is gather the data from the different sources, format the data on a way that it's easier to integrate with all the other time series (Data Collection and Integration). Then from a exploratory data analysis step, it's possible to get to know better the data and get some patterns and insights from it. This way an experiment can be designed and the data pre-processed, as each time series was collected in different ways, either in terms of record frequency, the nature of the sampling procedure, the precision of the measurement, or even the sensor or human error when recording the data, which will be more explained on chapter 3, to finally be given as input to model the toxin contamination event.

1.4 Contributions

This thesis contributed to the construction of a data pre-processing pipeline, and with the integration of the multiple time series collected from multiple sources, namely IPMA and Copernicus satellite program. MARETEC, a research center with its main activity focusing on water environment data extraction and modeling, also helped in providing data on water nutrients and hydrodynamic variables. Finally to get a model to identify the conditional dependencies between different variables, a Bayesian approach was adopted, and a framework already implemented for training the Dynamic Bayesian Network models was used, called MAESTRO, and it can be accessed in https://vascocandeias.github.io/maestro/.

Chapter 2

Related Work and Concepts

This chapter will be divided into two sections. A more theoretical one describing the background of the solution concepts on time series analysis and Bayesian networks parameters and structure learning. And a second one which brings the state of the art on the usage of Bayesian statistical models that already tried to solve this problem, by modeling the causality and even predicting the contamination event.

2.1 Background Concepts

2.1.1 Time Series and its Analysis

The source of information gathered to solve the shellfish contamination problem, is constituted by observed data collected in-situ from shellfish and water samples, by meteorological sensors or generated by theoretical models throughout time, named time series, which can record past patterns and information that can be used to forecast future behaviors or even find relation between other variables from an independent collection process.

Time series is a sequence of values of a random variable, measured along time, with equally time spaced intervals. Many exploratory tasks can extract information, and some indicative behavior from history including seasonal and trend behaviours, stationarity, correlation between time periods and variables, and forecasting. Seasonality is a component in which data patterns have stable frequency, i.e., recur over a fixed period, such as temperature measurements on summer periods. Identifying and interpreting this behavior can be important on this specific project, since variables like temperature and rainfall can show seasonal variations that could lead to phytoplankton growth and *E.coli* spread, and may be considered when modeling the time series. Trend is a component that shows an increasing or decreasing value in time series, extracted using a moving average on a window defined by the finer period identified.

An important concept on a time series analysis and modeling is stationarity, since the accuracy of a model will vary over time, i.e. sometimes could be right and some times do not. A time series is said to be stationary when the mean and variance are constant along time, so the variable distribution doesn't change with time. Seasonal behavior on time series is the antithesis of stationarity because it can result in a changing of variance over time, and trends can result in a varying mean over time [24]. 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. 2.1, is nothing less than the correlation or the degree of similarity between the current series and other lagged, usually being the Pearson coefficient.

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=1}^{T} (y_t - \overline{y})^2}. \tag{2.1}$$

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. 2.2:

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=k+1}^{T} (y_t - \overline{y})^2 \sum_{t=k+1}^{T} (y_{t-k} - \overline{y})^2}. \tag{2.2}$$

In this research context there are multiple variables observed over time, to be considered to predict future values of biotoxins concentration and HAB events, since each one may depend not only on its past values but also on other variable, defining what is a MTS. For MTS forecasting the dependence knowledge between each variables time series is important, and to calculate that the cross-correlation is measured between a pair of series.

The same method of lagging time series to calculate the correlation coefficient between two different time series, where 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 2.1: Two Different Time Series Lagged

Timestep	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

But 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 [25].

2.1.2 Bayesian Networks

Biological processes such as Eutrophication, specifically algal proliferation, and consequently shellfish contamination are complex and stochastic, so the probability representation seems appropriate given the uncertainty of the abiotic and biotic factors evolution.

Bayesian networks are a family of graphical representations of distributions, that uses a directed acyclic graph to map dependencies and independencies between multiple random variables, by setting one-way edges from the source to the target variable. These models are useful because both the structure and the parameters provide a natural representation for many types of real-world domains [26].

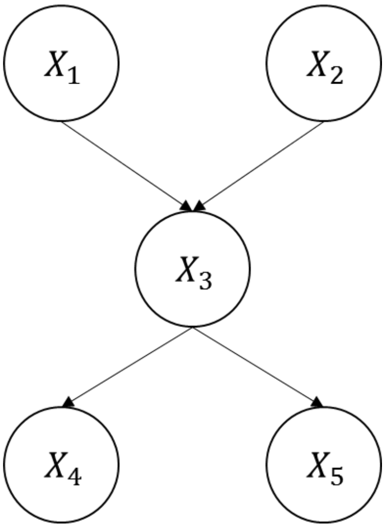


Figure 2.1: A directed acyclic graph of a set of random variables $X = X_1, X_2, X_3, X_4, X_5$.

This dependency graph allows to break a high dimensional distribution, into small factors and the factorization used on a Graphical Model such as a Bayesian Network (BN) already constrains all the possible probabilities over a set $X = \{X_1, \ldots, X_N\}$, which reduces the inference time complexity.

The factors $\phi = \{\varphi_1(X_1), \varphi_2(X_2), \varphi_3(X_1, X_2, X_3), \varphi_4(X_3, X_4), \varphi_5(X_3, X_5)\}$ in the Fig. 2.1, define all conditional dependencies and independencies, where each node X_i is associated with a conditional probability distribution (CPD), that turns into a marginal distribution on nodes with no parents like X_1 and X_2. A joint distribution of an event with evidence from variables X_i is then calculated applying a chain rule, that is a factor product of ϕ probabilities:

$$P\left(X_1, X_2, X_3, X_4, X_5\right) = P\left(X_1\right) \cdot P\left(X_2\right) \cdot P\left(X_3|X_1, X_2\right) \cdot P\left(X_4|X_3\right) \cdot P\left(X_5|X_3\right). \tag{2.3}$$

Which can be defined in a more generalized way, considering B the BN, and pa the node's parents:

$$P_B(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i|pa(X_i)). \tag{2.4}$$

There are two ways of getting a BN to model a problem, by empirically setting the structure G and the probabilities or parameters θ with expert knowledge, or by learning G as an optimization of θ.

$$\theta_{ijk} = P_B(X_i = x_{ik}|pa(X_i) = w_{ij}), \tag{2.5}$$

Where i is the index for all possible random variables on vector X, $k \in \{1, \ldots, r_i\}$ is the index for possible values or states r_i of a continuous or discrete variable X_i, respectively, $j \in \{1, \ldots, q_i\}$ is the index for possible combinations of values or states r_l on vector $q_i = \prod_{X_l \,\in\, pa(X_i)} r_l$ of the continuous or discrete parents' variables l of node X_i, respectively and θ_{ijk} is the parameter or the probability for a certain node X_i to take its configuration x_{ik}, given that its parents take their configuration w_{ij}.

On a collection or dataset D of records or instances with n random variables, a set of BNs B_n with different combinations of n variables is the search space to find which one maximizes the likelihood of the parameters θ_{B_n}. The heuristic used for this optimization problem its a commonly used scoring function, log-likelihood [27], and the search algorithm can be a simple hill-climbing. There are other approaches to get a better graph structure to represent D and at the same time a more generalized one, able to estimate a good approximation of the posterior of unobserved evidence.

After finding the optimal G, for modelling the dependencies of multiple time series variables, this framework can provide inferences of posterior probabilities, i.e. a conditional probability of given a set of observed evidence E, that are values from X_i variables, what is the most probable value of a time series variable X_i, $P(X_i = x_{ik}|X_j = e_i)$, $e_i \in E$. The inference is given by:

$$P(X_i = x_{ik}|E_i = e_i) = \frac{P(X_i = x_{ik}, E_i = e_i)}{P(X_j = e_i)}. \tag{2.6}$$

A structure restriction algorithm such as Belief Propagation, is usually used to perform an exact inference, since the joint probability calculated with factorization using Eq. 2.4 can be a complex task, thus tree networks are adopted or even approximation inferences in clustered graphs.

It has already been showed that the biomass of an harmful algae named Microcystis were mainly explained by a combination of abiotic factors like water temperature and total nitrogen, within others [28]. But regarding the cyanobacteria process of releasing the toxin microcystin (lysis), it isn't an immediate outcome, i.e. the presence algae bloom predicted, may not mean that in that same time instance there will be high concentration of biotoxins, so the abiotic and biotic factors should not be modelled as synchronized variables but modelled with different time-dependence. And such framework is built with Dynamic Bayesian Networks (DBNs) [28].

2.1.3 Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBN) are an extension of BNs, 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:

$$P\left(X^{(0:T)}\right) = P\left(X^{(0)}\right) \prod_{t=0}^{T-1} P\left(X^{(t+1)} \middle| X^{(0:t)}\right) \tag{2.7}$$

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. (2.7):

$$P\left(X^{(0:T)}\right) = P\left(X^{(0)}\right) \prod_{t=0}^{T-1} P\left(X^{(t+1)} \middle| X^{(t-m:t)}\right) \tag{2.8}$$

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.

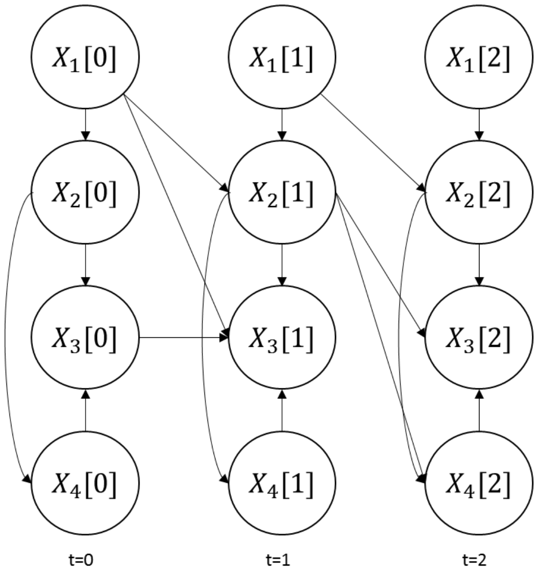


Figure 2.2: DBN unrolled for 3 time steps.

A fixed template transition model such as in fig. 2.2, 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.

Inference in DBNs estimates values of unobserved variables, at a given time step, using algorithms like forward-backward or particle filtering. If the goal is predicting the value in the past, it is known as Smoothing. If it is for predicting the values of variables with no evidence at the current time, it is known as Filtering and in future time slices is considered Prediction.

2.1.3.A Learning a Stationary Model Structure and Parameters

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, as in fig. 2.3. Then rather than using mutual information to calculate the score for each edge, as in [29] 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

$$e_{ij} = s_{ij} - s_i, \tag{2.9}$$

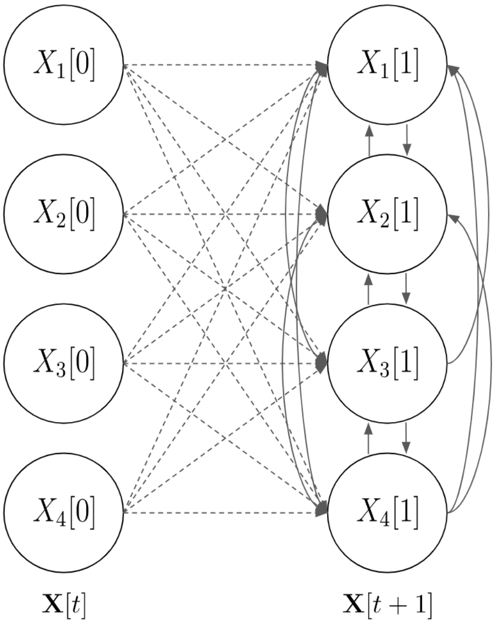


Figure 2.3: Complete DAG with all the possible connections.

with the s_i being calculated as

$$s_i = \max_{X_p[t] \epsilon X[t]} \phi i(X_p[t], D_t^{t+1}), \tag{2.10}$$

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

$$s_{ij} = \max_{X_p[t] \epsilon X[t]} \phi i(X_p[t] \cup X_j[t+1], D_t^{t+1}). \tag{2.11}$$

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, $MDL(G|D) = LL(G|D) - \frac{1}{2}log(||D||).||G||$. 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. 2.4.

The $ENT_D(X_i[t+1]|X_p[t])$, 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

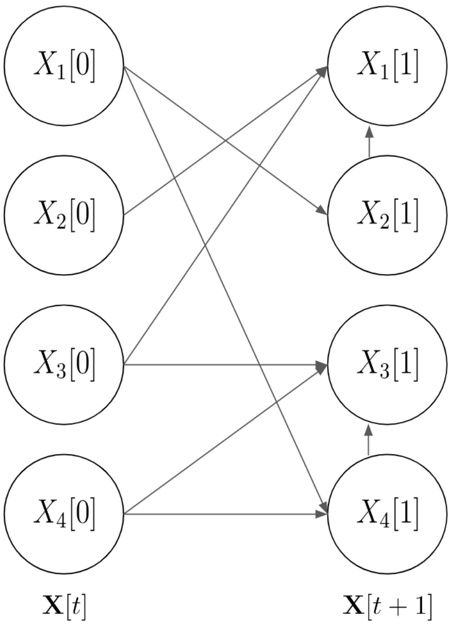


Figure 2.4: Learned DBN.

$$ENT_D(X_i[t+1]|X_p[t]) = \sum_{i=1}^{n}\sum_{j=1}^{q} P(n_i,q_j) log \frac{P(n_i,q_j)}{P(n_i)P(q_j)}. \tag{2.12}$$

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.

The next section will revise the state of the art of Bayesian statistical models for shellfish contamination phenomenon explanation but mostly for forecasting.

2.2 State of the Art

2.2.1 Bayesian Statistical Models on Shellfish Contamination

There has been studies focused on DBN framework to model different type of problems like for water eutrophication factors causality and inference [8, 30], and others for emerging HABs risk or biotoxins concentration like [31].

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 [31].

There is also a study that take the advantage of modelling multivariate time series with a DBN, and from the transition networks on intra and inter time-slices it can extract outliers, i.e. a set of observed evidence with low probability, and can be used to help identifying anomalous combination of values on the biotic and abiotic factors of the MTS available, that could mean a contamination moment on this particular problem [32].

It could be relevant as well, to map the causality and inter-relations between contiguous areas, by applying an hierarquical BN representation [31], of coastal areas with multiple sampling points like estuaries or lagoons, and even between some of the bigger areas.

2.2.2 Shellfish Contamination Biological Theory

Normally marine biotoxins concentration are extracted from shellfish tissue samples, since it is the most direct method of monitoring shellfish contamination, which can be a process that differs from species to species, since the accumulated amount of toxins within the tissues of filter feeding shellfish, may not be that evidential since each specie has a proper accumulation and elimination process. The phytoplankton concentration is measured on a specific instance in time, and the biotoxins concentration at toxic levels may only be expressed some days after an HAB event, which may vary again because of the different species metabolism. The increase on proliferation and density of phytoplankton is studied along many studies, that try to prove which are the most indicative abiotic factors that contribute to this phenomenon [33].

Chapter 3

Proposed Solution and Methodology

3.1 Data Collection and Integration

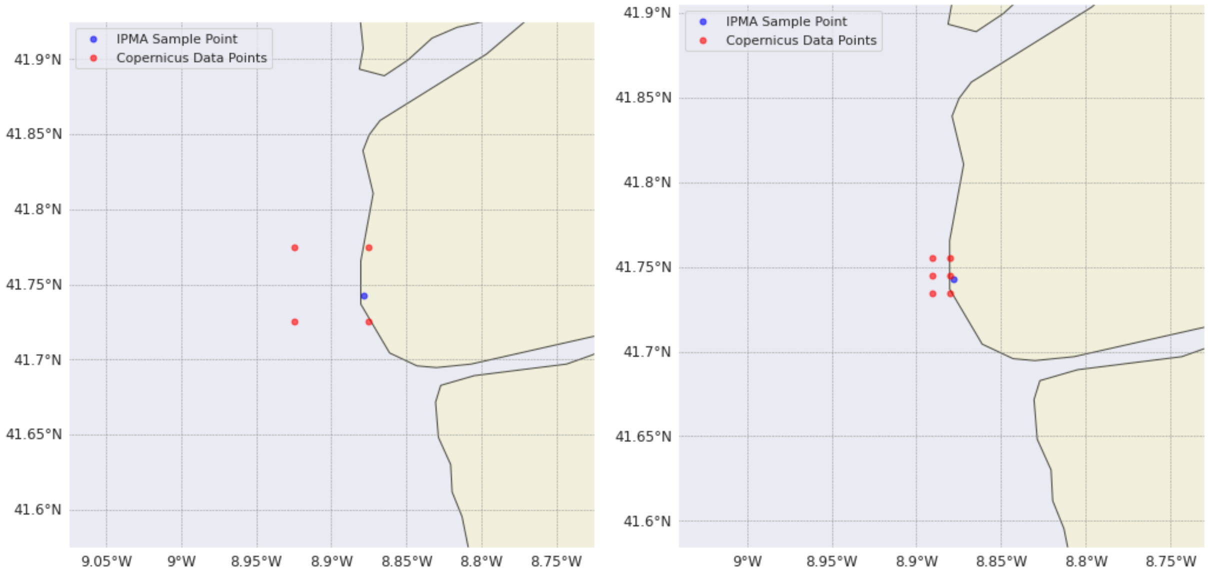
3.1.1 IPMA In-situ and Meteorological Data

IPMA's biological time series, of DSP, ASP and PSP both toxins and phytoplankton concentration was collected from IPMA's website monthly report. The toxins records are ordered by date on a table that registers its concentration on each sampling station, but apart from having to correct some production areas, sampling stations and species names, some sampling tests values that are indicative of some thresholds like the value being above 2400 µg/kg, or non-detectable (ND), non-quantifiable (NQ) or even not registered (NR), were replaced by the threshold values or missing value respectively. For example the values registered as ">2400" were replaced by 2400, the "ND" or "NQ" values were replaced by the toxin quantification detectable limit, as being 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". 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 environmental factors collected daily by meteorological stations, were provided by IPMA from a set of 20 stations with time series of air temperature, wind intensity and direction, and precipitation as daily averages values from 2015-2020. These time series records only had a value of -990 each time there were a missing value, that were replaced by "n/a" for a posterior analysis. From 2021 the data was provided with a hourly granularity, which needs to be aggregated by day, but before replace the -990 value, representing the missing value for "n/a", and then an average aggregation function can be applied by day except for wind direction that a mode function is preferable. Then to integrate with the sampling point granularity, the meteorological and sampling stations were merged by the closest locations, originating a meteorological dataset for each sampling point.

3.1.2 Copernicus Remote Sensing Data and MARETEC Data from MOHID Model

Through Copernicus satellite remote imagery of Atlantic Ocean regions it was collected scientific photography of the nearshore points closest to IPMA's sampling points. This satellite images get information through recording the electromagnetic spectrum wavelengths on each pixel, of sea surface temperature (SST) and clorophyll-a (CHL-A). Each product receives as input a set of geographic coordinates within an area covering a sampling station geolocalization, and then it outputs a multidimensional array with indices composed by time instance, latitude and longitude and the respective value. So as the multidimensional arrays varied in terms of granularity, the indexing passed through a process of finding the best latitude and longitude values around the sampling stations coordinates, by visualizing on a map like the one in fig. 3.1 for each production area.



(a) Analyzed SST at L1.

(b) Analyzed CHL-A at L1.

Figure 3.1: Copernicus Data Points around the sampling stattion of L1

Finally the MARETEC estimated water property and hydrodynamic variables from MOHID model, for some production areas, and the original files on an ets format were transformed to a csv format.

3.1.3 Merge Multivariate Time Series

There are 38 production areas across the Portuguese coast, some having more than one sampling station, totalling 60 sampling points, where each one could have different species, which results on a set of 263 entities with the respective key as (Production Area — Sampling Point — Specie). From these 263 entities a preliminary selection was applied to only filter in entities with at least 156 records of biotoxins concentration, i.e. 3 years of weekly records of the variables of interest, resulting on only 33 entities, almost one per production area, from what could be the most relevant specie within each production area.

Now from these preliminary entities IPMA's biological and meteorological data is merged. IPMA's biological data is collected in-situ, the biological properties on water samples, like toxic phytoplankton for each production area and the biotoxins concentration from shellfish tissue samples on the respective sampling point of that production area. Both sets of IPMA's biological records have a weekly granularity, but they don't match neither on the day of the week nor the weekly frequency, since there are some missing records during some periods, and also some sampling analysis can get repeated on the same day or happen two times a week on different days. So a new weekly frequency index is created for both time series, from 2015/01/01 to the 2020/12/31, 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. In case it hasn't any record on that week period, it means that no sample record where registered around that time, so it is a real weekly missing record. Since IPMA's meteorological data has a daily granularity, to merge it is only necessary to get the day closest to the day of the new index, except for wind direction that will use a rolling mode to get the weekly predominant direction, then it is a basic merge by date.

To merge the previous multivariate time series with the Copernicus remote sensing imagery data of clorophyll-a, sea surface temperature, and the MARETEC data, both were merged by getting the record on the day closest to the day of the new index.

3.2 Exploratory Data Analysis

The data analysis allows to extract data patterns and get insights from what variables could trigger the natural chain of biological effects and generate a contamination process on shellfish. One approach is to get evidence of some relation between variables, like analyzing two time series against each other throughout the years and with the time series lineplots try to identify same trends with consistent positive or negative relations. After that confirm through correlation and cross-correlation if that trend direction in fact results on a high correlation, and finally prove it by getting an inter or intra slice relation within the Bayesian network.

The study started by finding the main affected areas, that on a more deep analysis could be labeled as the toxic contamination events main area indicators and/or the more proliferating ones. After that the study cases were defined within different coastal zones, to investigate a possible dependency chain between the concentration of phytoplankton and biotoxins of a relevant area and the concentration of its neighboring areas. This case studies definition required as well the same process of finding a common period with the minimal requirements of getting at least a time series with length of 104 records (2 years), both having less than 20% missing values but now at least one of the entities should have 5% of high toxin concentration and at least one entity should have 5% of high phytoplankton concentration. If there isn't a common period that respect this requirements, that pair of areas or set of areas are no longer considered for analysis.

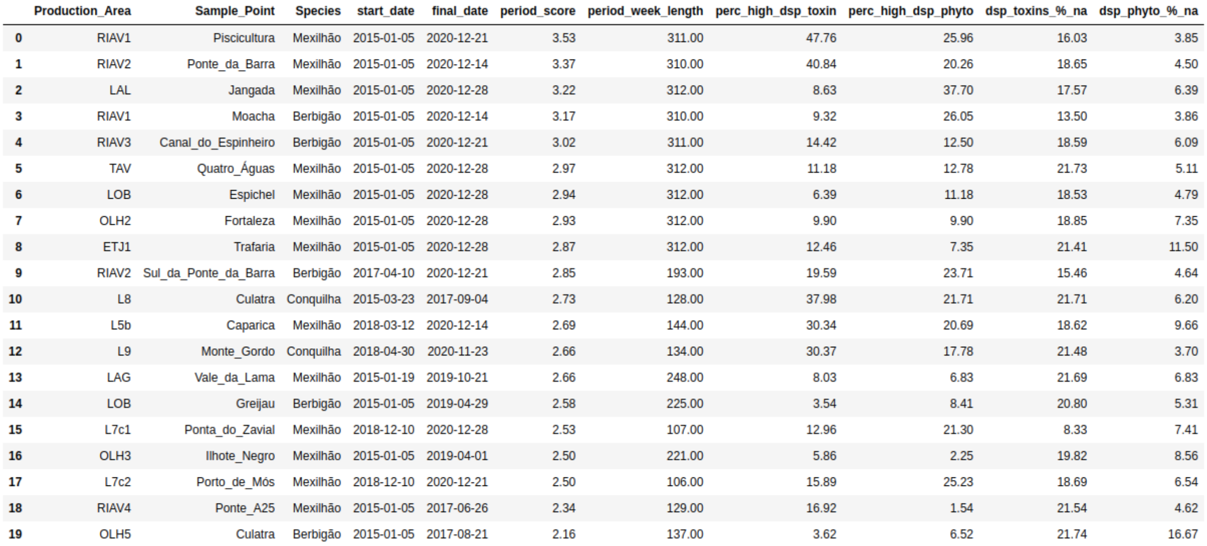
That investigation began with a bivariate analysis between time series of toxins on an area against toxins on a contiguous area, between the phytoplankton time series, by calculating the cross-correlation between each pair of areas within each coastal zone case study, with the purpose of getting a preliminary notion of inter and intra-timeslice relations I could expect. The stronger and more direct relations can easily be revealed from this analysis, paying more attention to correlations on when the time series are lagged by 1 or 2 time steps, as expected in theory. But some relations may need other complementary variables to explain the toxic phenomenon, which may be found when modeling the data with the dynamic Bayesian networks.

3.2.1 The case study area and specie

Since there are 33 sampling stations, the problem scope is too much complex to study all the relations between all the variables and all areas. So an experiment must be designed to reduce this complexity and get multiple simpler contamination events to solve and explain the causality in the beginning.

3.2.1.A Production Area Analysis

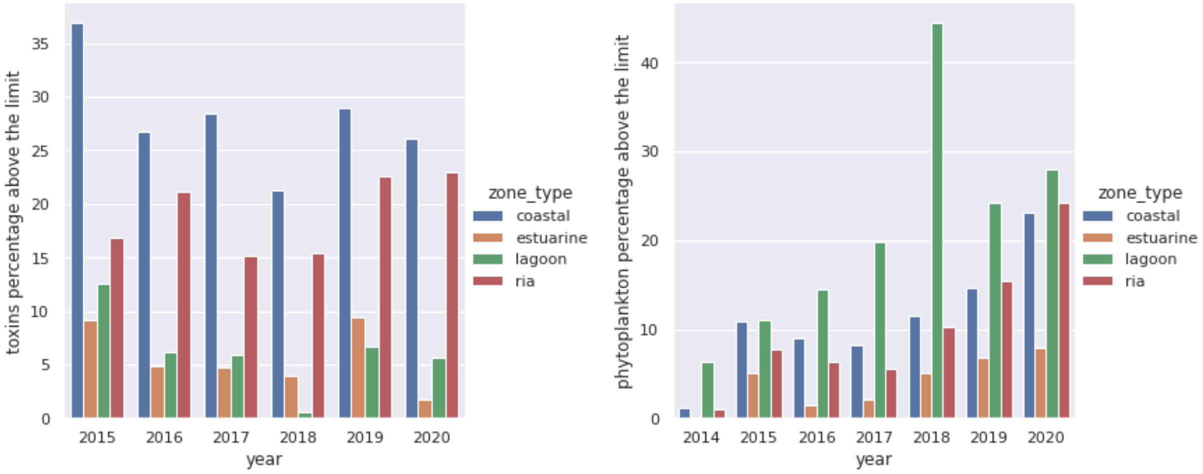
To help on reducing the problem scope a method of ranking the entities by its time series high toxicity and completeness was designed. First of all iterate over all 33 entities and rank every possible period of each entity from the biggest one starting on the first possible date and ending in the last record date to the smallest one possible with a minimum of 104 records, i.e. 2 years of records, by iterating the starting date and then iterate over all dates until the last record date to define the final date of the period. The entity period scoring function takes into account the percentage of high toxicity events, the percentage of missing values of both toxins and phytoplankton, the number of weekly records and also the percentage of high phytoplankton concentration ($perc\_high\_dsp\_toxin$, $dsp\_toxins\_\%\_na$, $period\_week\_length$ and $perc\_high\_dsp\_phyto$, respectively, at the table in fig. 3.2). The number of weekly records, and both percentages of high toxicity and phytoplankton concentration are converted to a range of values between 0 and 1, where the closer to 1 the bigger the number of records or percentage, whereas with missing values percentage it's the other way around with the same range from 0 to 1 but now the closer to 1 the smaller the percentage, which is what is desired, and in the end all is summed up as a final score. The idea is to get for each area the biggest and most complete time series period in terms of the primary variables of interest (biotoxins and phytoplankton), with less than 20% of missing values, but also with a minimum percentage of high toxicity events of 5%, 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) since this is the event of main interest on this study. Additionally even if there aren't periods on an entity that fills the high toxicity percentage conditions, there is still the possibility of being relevant to model the causality of a contamination event on its neighboring areas and be a potential contamination dispersion point, if it gets a percentage of phytoplankton concentration higher than 5%, which means getting more than 200 cel/L. In the case when the entity gets a period with the minimum considered high toxicity events and phytoplankton concentration it means that most probably is a type of production area prone to proliferate HAB blooms and the shellfish metabolic capacity of accumulating biotoxins is higher than the elimination. When it only registers more than 5% of high toxicity events but not high phytoplankton concentration, could mean that is a type of production area that doesn't retains so much toxic phytoplankton but it is sufficient to contaminate the shellfish with toxins.



	Production_Area	Sample_Point	Species	start_date	final_date	period_score	period_week_length	perc_high_dsp_toxin	perc_high_dsp_phyto	dsp_toxins_%_na	dsp_phyto_%_na
0	RIAV1	Piscicultura	Mexilhão	2015-01-05	2020-12-21	3.53	311.00	47.76	25.96	16.03	3.85
1	RIAV2	Ponte_da_Barra	Mexilhão	2015-01-05	2020-12-14	3.37	310.00	40.84	20.26	18.65	4.50
2	LAL	Jangada	Mexilhão	2015-01-05	2020-12-28	3.22	312.00	8.63	37.70	17.57	6.39
3	RIAV1	Moacha	Berbigão	2015-01-05	2020-12-14	3.17	310.00	9.32	26.05	13.50	3.86
4	RIAV3	Canal_do_Espinheiro	Berbigão	2015-01-05	2020-12-21	3.02	311.00	14.42	12.50	18.59	6.09
5	TAV	Quatro_Águas	Mexilhão	2015-01-05	2020-12-28	2.97	312.00	11.18	12.78	21.73	5.11
6	LOB	Espichel	Mexilhão	2015-01-05	2020-12-28	2.94	312.00	6.39	11.18	18.53	4.79
7	OLH2	Fortaleza	Mexilhão	2015-01-05	2020-12-28	2.93	312.00	9.90	9.90	18.85	7.35
8	ETJ1	Trafaria	Mexilhão	2015-01-05	2020-12-28	2.87	312.00	12.46	7.35	21.41	11.50
9	RIAV2	Sul_da_Ponte_da_Barra	Berbigão	2017-04-10	2020-12-21	2.85	193.00	19.59	23.71	15.46	4.64
10	L8	Culatra	Conquilha	2015-03-23	2017-09-04	2.73	128.00	37.98	21.71	21.71	6.20
11	L5b	Caparica	Mexilhão	2018-03-12	2020-12-14	2.69	144.00	30.34	20.69	18.62	9.66
12	L9	Monte_Gordo	Conquilha	2018-04-30	2020-11-23	2.66	134.00	30.37	17.78	21.48	3.70
13	LAG	Vale_da_Lama	Mexilhão	2015-01-19	2019-10-21	2.66	248.00	8.03	6.83	21.69	6.83
14	LOB	Greijau	Berbigão	2015-01-05	2019-04-29	2.58	225.00	3.54	8.41	20.80	5.31
15	L7c1	Ponta_do_Zavial	Mexilhão	2018-12-10	2020-12-28	2.53	107.00	12.96	21.30	8.33	7.41
16	OLH3	Ilhote_Negro	Mexilhão	2015-01-05	2019-04-01	2.50	221.00	5.86	2.25	19.82	8.56
17	L7c2	Porto_de_Mós	Mexilhão	2018-12-10	2020-12-21	2.50	106.00	15.89	25.23	18.69	6.54
18	RIAV4	Ponte_A25	Mexilhão	2015-01-05	2017-06-26	2.34	129.00	16.92	1.54	21.54	4.62
19	OLH5	Culatra	Berbigão	2015-01-05	2017-08-21	2.16	137.00	3.62	6.52	21.74	16.67

Figure 3.2: Entities Scoring and Statistics Table

A ranking table with all the top entities was ordered by the score. By looking at the table in fig. 3.2 and according to the previous knowledge on shellfish market predominant production areas, 3 different testing scenarios, with the potential combinations of areas around the top entities in different zones, as the Ria de Aveiro (RIAV1), Lagoa de Albufeira (LAL) and Quatro Águas at Ria Formosa (TAV), could be set. By defining 3 scenarios with closer neighboring entities, I could model the probable contamination chain around the top area, which can be the most affected by toxin contamination by the surrounding areas and/or gather good oceonographic conditions to hold and proliferate phytoplankton in a way that is the responsible for contamination events on its neighbors. As it was expected more to the North, Ria de Aveiro (RIAV1), more precisely at Piscicultura, is one of the areas that got the most informative length of records within the minimum requisites of relevancy defined previously, and showed that is both informative on toxicity and phytoplankton agglomeration. I would expect that by the proximity between the areas to model the contamination causality the neighbors would be LOB and L2, but unfortunately there weren't considered relevant in terms of toxicity and completeness. More to the center the selected top area was Lagoa de Albufeira (LAL), Jangada as the sampling point, since its a lagoon and it shows the same conditions of phytoplankton retention as RIAV1 which can be confirmed by looking at the graphs on Figure 3.3 of the frequency of high toxicity on shellfish and high HAB toxic events among the different type of regions, that lagoons and rias are most likely to register high toxic phytoplankton than high toxicity on shellfish. And LAL potential neighbors are Trafaria (ETJ1) and Caparica (L5b) by proximity, which also are among the relevant areas. At the South coast closer to the windward side the LAG as the top area, and more to the leeward side TAV seems to be a top area, i.e. the most affected area on that side, where the neighbors can be LAG and other top areas around as well, like Fortaleza (OLH2) and Monte Gordo (L9).



(a) High toxicity frequency distribution over the years. **(b)** High HAB frequency distribution over the years.

Figure 3.3: Frequency of Toxic events per type of region over the years.

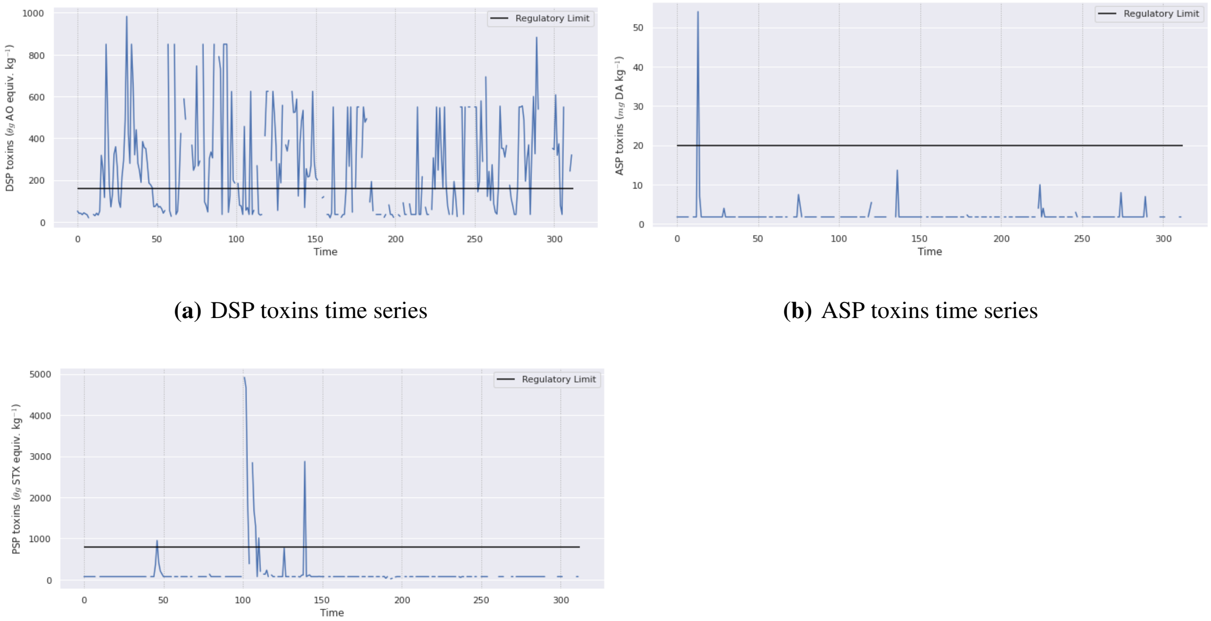
When assuming that my areas relations are only pairs, maybe I'm loosing some relevant information since the phytoplankton can go two different ways. So the preliminary models, when I had no idea of the contamination impact between areas, I can have 2 neighbors from both sides of the local top area. In the end to keep the model simpler with less complex structures and parameters, at least the first ones, it's easier and feasible to do it in pairs around the top area.

On a forecasting perspective the selection of the best period as being the part of the time series with more events of interest, could be seen as a biased selection, by getting the model to overfit on a specific time period and most certainly getting worse results when a new year records with less events of interest is inserted in the model. But on a dependency study point of view, mainly when using Bayesian models, if the prior knowledge on the phenomenon is richer it converges easily to a better dependency structure.

3.2.1.B Specie and Toxin Analysis

Regarding the different species and toxins present on different sampling points, this project focused on the toxin that showed more frequency on values above the regulatory limit, as the DSP dominance is apparent on the time series plots in Figure 3.4, being these toxins the most common in the Portuguese coast [34].

And the mussels as the specie that were considered the most indicative of a contamination event on the production area, since is the reference specie to alert a high toxicity levels of this toxin type, because they have the better biotoxin accumulation capacity than elimination [12, 35].



(a) DSP toxins time series

(b) ASP toxins time series

(c) PSP toxins time series

Figure 3.4: Toxins Time series of Ria de Aveiro, Piscicultura

3.2.2 Cross-Correlation Analysis

The cross-correlation analysis were done using both Pearson, to find linear relations between variables, and the DPCCA already explained in the previous chapter 2, for all the areas identified as relevant. By only looking at the cross-correlation plots, if the correlation with lag 1 or 2 are not yet 0, there is already some indication that the past observations may help to predict the present value so it is an important indicator of some relation between time series.

3.3 Data Pre-Processing and Feature Selection

To find model parameters and Bayesian network structure, and train it, some data requirements already described on chapter 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.

For simplification purposes, but maintaining the main goal of modeling the contamination impact between different production areas, the time series that will be used from each area are only the DSP toxins and DSP phytoplankton.

3.3.1 Best Period Selection

The best period selection assures that I can reduce the bias that will come from filling missing values with synthetic data. The process is the same as the one that ranks each single area in terms of time series completeness and high toxicity frequency, but now the period interval iterated is applied to both areas of each pair, as well as the conditions to be considered a common good period. The best one is chosen by looking at the one that got the best score among all, which reduce the number of records of both areas time series. The idea behind obtaining the best period, is also to obtain richer time series with the purpose to provide more toxic events to the DBN models to better model the dependencies.

3.3.2 Imputation

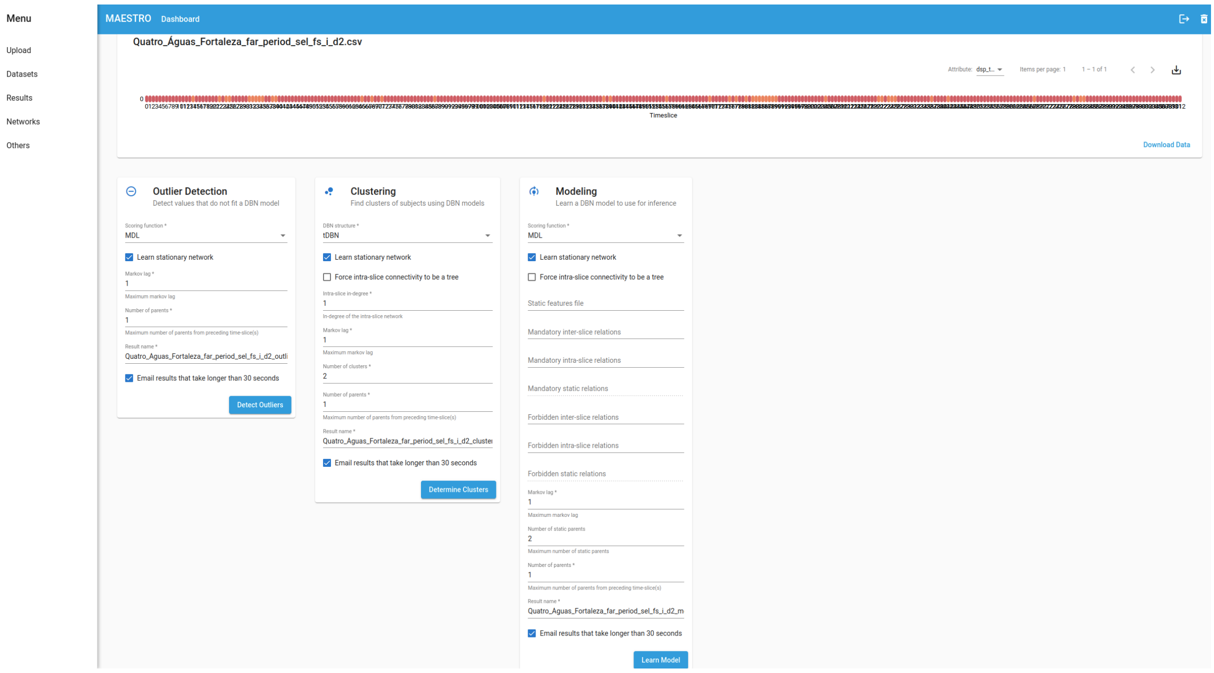
The missing values from phytoplankton and mostly toxins may come from the fact that there are no need for sampling in-situ during the period that the area is closed, since there are some rules of thumb for how much weeks it should be closed. After the record selection a filling strategy was divided in two methods. The first one being the moving average, which only completes the time series records that have at least two complete records values on a window with a length of 7 records, where the one in the center is the missing record, so it looks for the average value from 3 weeks before to 3 weeks after. But this method cannot fill all missing records so a more conservative is then used to fully fill the DSP toxin and phytoplankton time series, with their default value of detectable limit.

3.3.3 Discretization

A first reduction on the levels of discretization per variable could be the most coarser one, with two levels of discretization for both toxins and phytoplankton time series, which would be below the toxic threshold and above, but the the main adapted one was chosen manually with some theoretical intuition, and by following a study that got better Bayesian model results when having more than just two levels, but less than five [36]. The DSP toxin time serie 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. 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

The idea is to study the impact each area has on other areas when its phytoplankton and toxin concentration are at their highest levels or near the limit a week or two weeks before their neighbors also observes that toxic proliferation increase. The graphical model web tool used to obtain the dynamic bayesian network to model this contamination dependency scenarios was MAESTRO, see fig. 3.5. MAESTRO includes data pre-processing tools with imputation and discretization methods to prepare any dataset variable to be a valid input to train a DBN model, but as I wanted to make a more specific data treatment to this particular time series, the data was already prepared before being loaded to the MAESTRO platform. Since all the time series were already processed, a DBN model configuration page appears automatically with multiple ways to choose the model learning process parameters.



Quatro_Águas_Fortaleza_far_period_sel_fs_i_d2.csv

Attribute: dsp_t...   Items per page: 1   1 – 1 of 1

Outlier Detection
Detect values that do not fit a DBN model

Scoring function *: MDL
☑ Learn stationary network
Markov lag *: 1 (Maximum markov lag)
Number of parents *: 1 (Maximum number of parents from preceding time-slice(s))
Result name *: Quatro_Aguas_Fortaleza_far_period_sel_fs_i_d2_outli
☑ Email results that take longer than 30 seconds

Clustering
Find clusters of subjects using DBN models

DBN structure *: tDBN
☑ Learn stationary network
☐ Force intra-slice connectivity to be a tree
Intra-slice in-degree *: 1 (In-degree of the intra-slice network)
Markov lag *: 1 (Maximum markov lag)
Number of clusters *: 2
Number of parents *: 1 (Maximum number of parents from preceding time-slice(s))
Result name *: Quatro_Aguas_Fortaleza_far_period_sel_fs_i_d2_cluste
☑ Email results that take longer than 30 seconds

Modeling
Learn a DBN model to use for inference

Scoring function *: MDL
☑ Learn stationary network
☐ Force intra-slice connectivity to be a tree
Static features file
Mandatory inter-slice relations
Mandatory intra-slice relations
Mandatory static relations
Forbidden inter-slice relations
Forbidden intra-slice relations
Forbidden static relations
Markov lag *: 1 (Maximum markov lag)
Number of static parents: 2 (Maximum number of static parents)
Number of parents *: 1 (Maximum number of parents from preceding time-slice(s))
Result name *: Quatro_Aguas_Fortaleza_far_period_sel_fs_i_d2_m
☑ Email results that take longer than 30 seconds

Figure 3.5: MAESTRO configuration page for training a 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, that empirically should be of 1 or 2 markov lags, 1 or 2 weeks respectively, since assuming there is a phytoplankton dispersion period before the shellfish accumulating the toxins and then after another week period there may be exist evidence of toxin contamination. 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 one variable may not be able to fully explain the behavior 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. Intuitively the more independent variables are modeled as parents of a possible toxin contamination, the more the conditional probability of having a high dsp toxin concentration knowing the different combinations of that independent variables decreases. So the number of maximum parents must be limited. And to restrict the search space on the optimization process of learning the best DBN structure forbidden inter-slice or intra-slice relations may be defined to help focus on the more informative interdependencies. In this case, 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, but assuming this is an empirical filter this must be optimized in future work.

Chapter 4

Experimental Results and Discussion

In this chapter, the results obtained were focused on the coastal zone where there are more entities within the preconditions tailored for this experiment, settled and explained in the previous Chapter, such as Vale da Lama (LAG), Fortaleza (OLH2), Quatro Águas (TAV) and Monte Gordo (L9). This set of production areas are also promising to show the hypothetical predictive linkage of an area with its neighbors because of its distance between each other and the geographical 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 [37].

4.1 Cross-Correlation Results

To calculate the correlation between entities DSP toxin and DSP phytoplankton time series, the cross-correlation methods used were the traditional Pearson method and DPCCA, already presented, as another method more appropriate to evaluate the correlation between time series without the influence of trend and other lags.

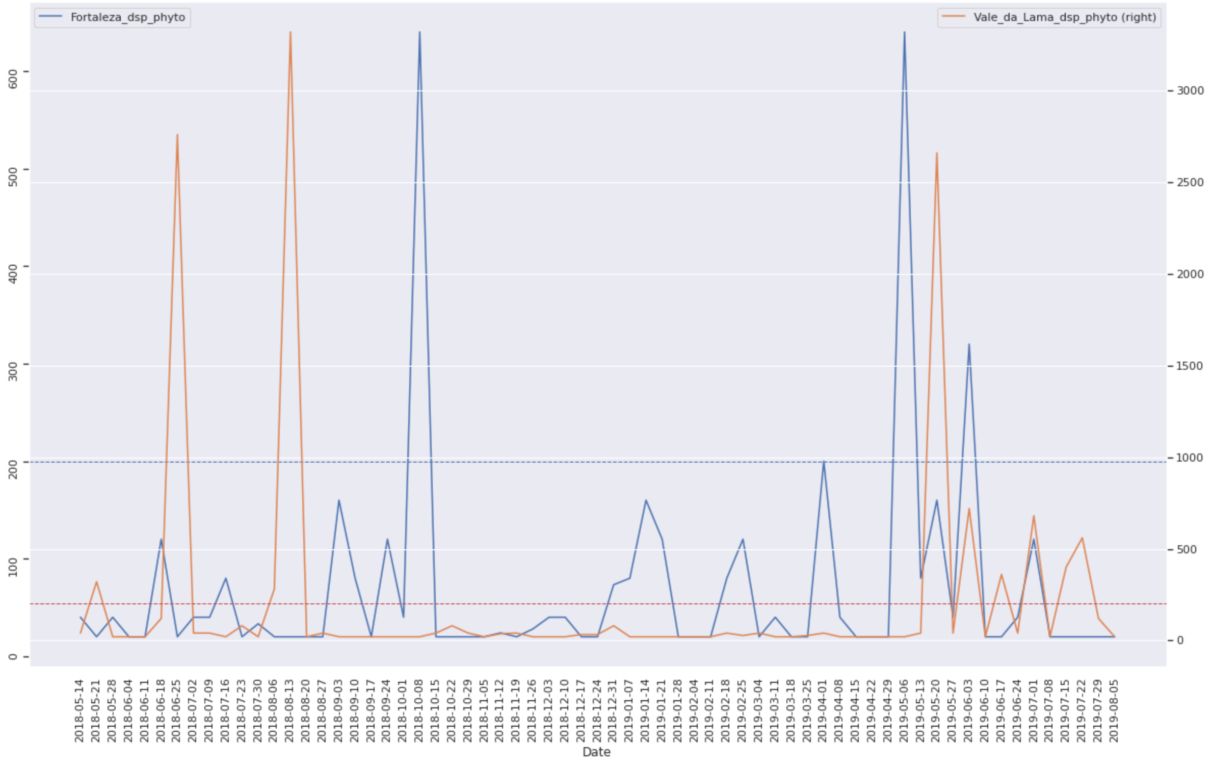


Figure 4.1: Fortaleza (OLH2) and Vale da Lama (LAG) phytoplankton time series on a specific period

For every possible pair within the case study scenario, the correlation was calculated for all the multiple pairs of time series, only considering the DSP toxin and phytoplankton time series for each entity. First of all 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 of the lagged time-window settled between -8 and 8, which is what is more important for our study on the impact between entities on distant areas to have a predictive perspective for a longer term notice.

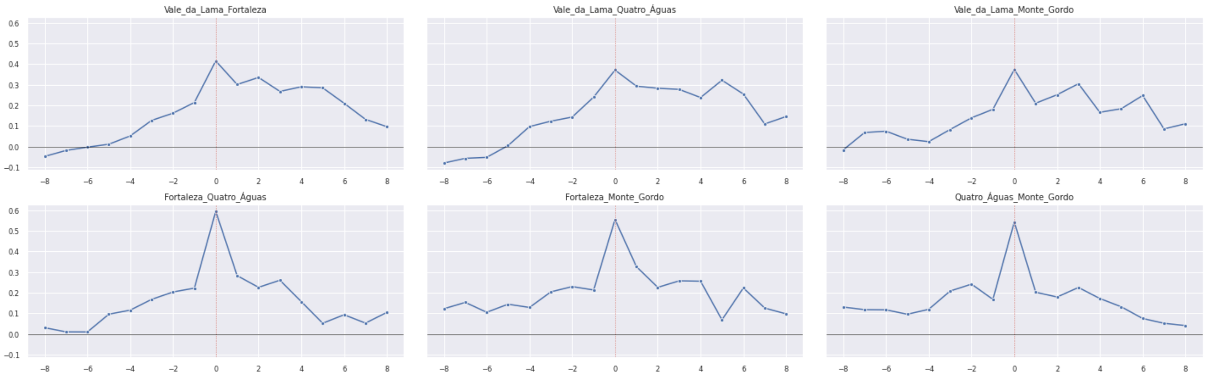


Figure 4.2: South Case Study Toxins-Toxins Pearson Correlations

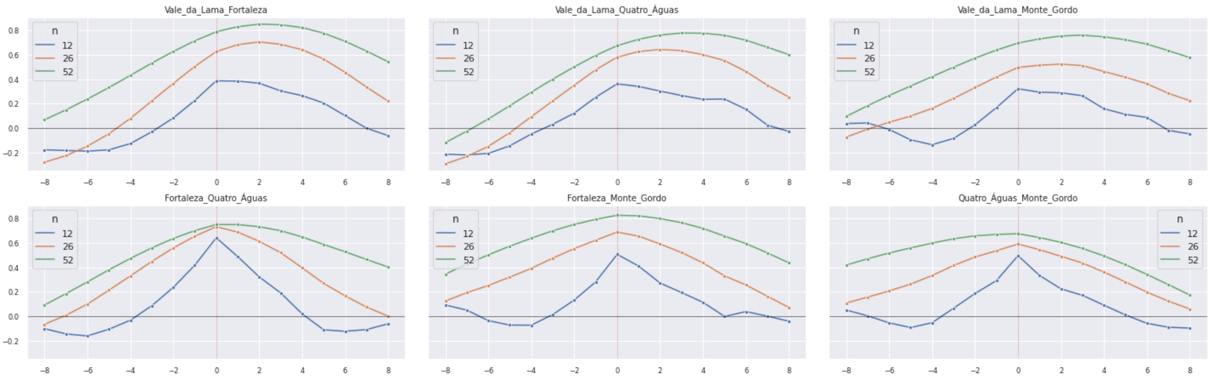


Figure 4.3: South Case Study Toxins-Toxins DPCCA Correlations

Now analyzing the direction of the correlation in terms of time series lag for each pair, all the phytoplankton-phytoplankton, in fig. 4.4 and fig. 4.5, correlations between LAG and the other areas are specially highly positive on the lag -1. By looking at a specific period of both phytoplankton time series at Vale da Lama and Fortaleza in the lineplot in fig. 4.1, some consecutive relation of high concentrations can be visualized, but there is not total certainty of that dependency direction dominance. Also the correlation between DSP toxins at Quatro Águas and Fortaleza DSP toxins apart from having a lower Pearson correlation on both lag -1 and 1, the DPCCA reveals some relation for both lag direction, which means both areas could have a high positive relation between each other in terms of toxins concentration, fig. 4.2 and fig. 4.3, on the same timeslice or on lag -1 where the DSP toxin concentration at TAV is somehow indicative of the DSP toxin growth in OLH2, and the opposite on lag +1. This 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. Of course this assumption is being made not taking into consideration, the different geographic and maritime conditions surrounding the areas.

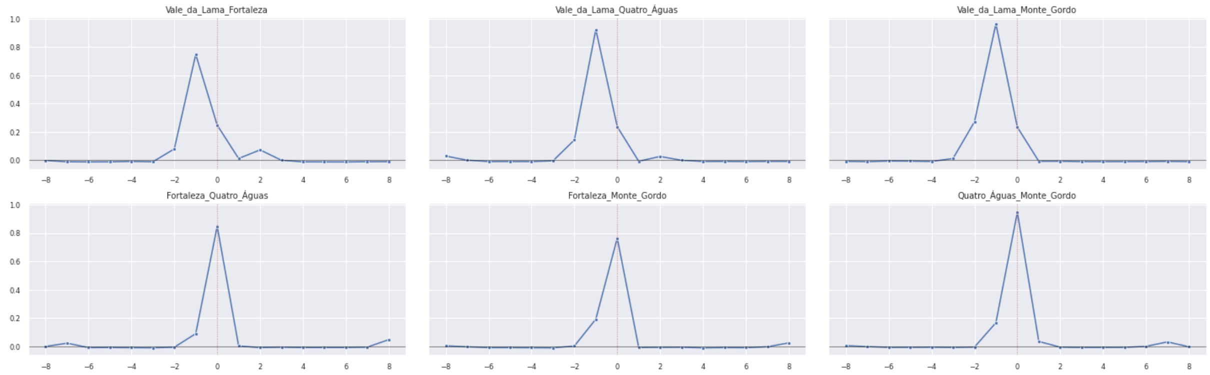


Figure 4.4: South Case Study Phyto-Phyto Pearson Correlations

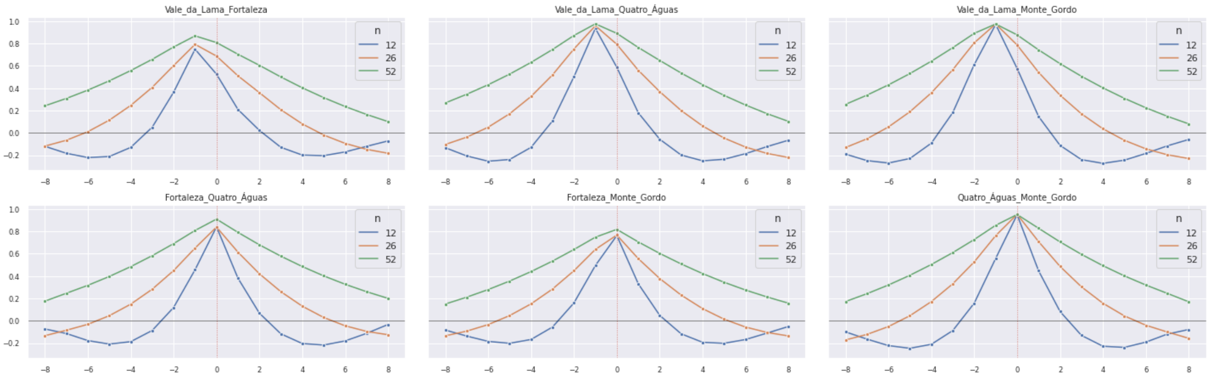


Figure 4.5: South Case Study Phyto-Phyto DPCCA Correlations

There is a consistency in terms of lagged direction with the highest correlation between every time series pair (phytoplankton-phytoplankton and toxins-toxins), which is from West to East for phytoplankton and the opposite direction for toxin impact as it is observed in fig. 4.6.

A high correlation coefficient, on this case a correlation closer to 1 or -1, does not mean a certain dependency relation between both random variables, and there is even less certainty if the correlation is around values that merely indicate some degree of positive correlation like 0.4 or 0.3 as the most lags above get. So a way to study a more multivariate dependency analysis is to model this pairs of variables for each entities pair with the Bayesian models, and find which exogeneous variables, i.e. variables that are not explained by other variables within a model, best help provide information, explain and predict each endogeneous variable, i.e. variables that are explained by other variables within a model, in the next weekly timeslice.

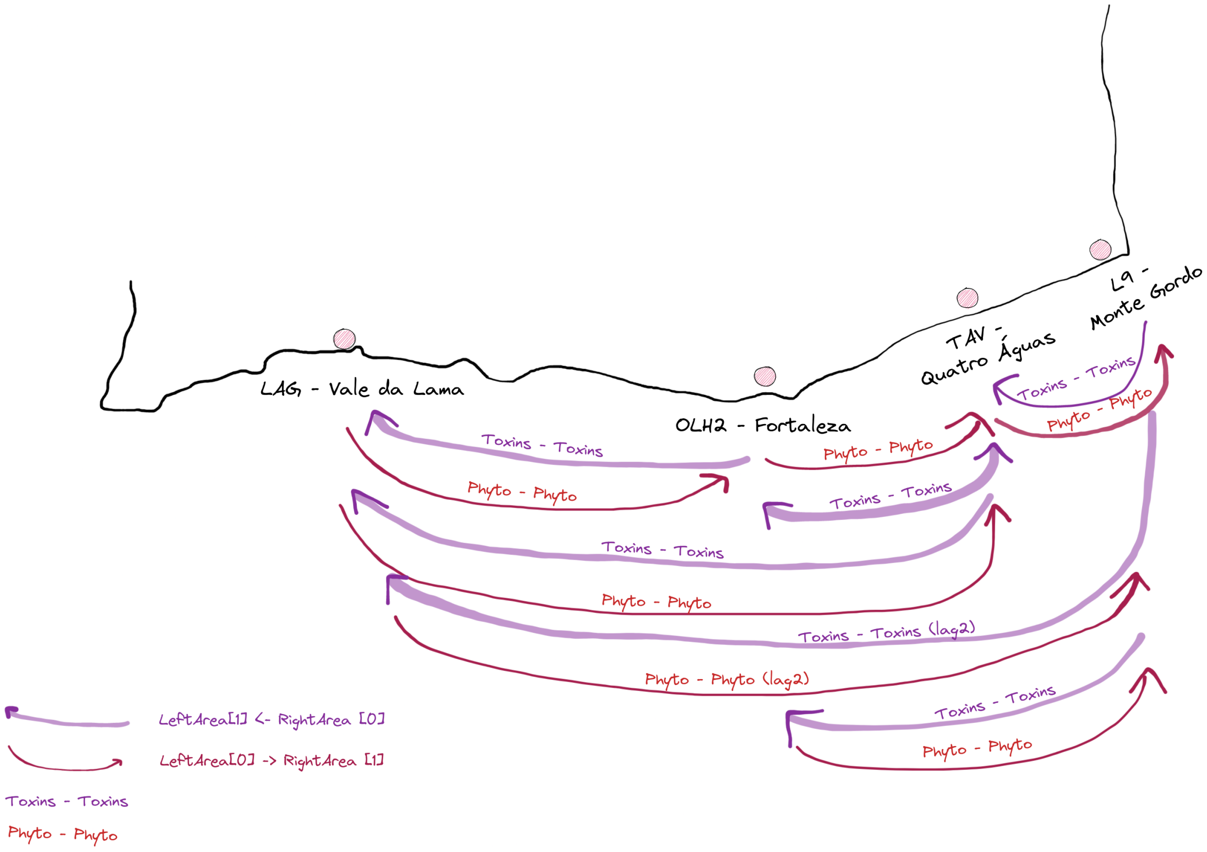


Figure 4.6: Map of South Areas and their interdependencies only based on cross-correlations

4.2 Dynamic Bayesian Models Results

Each pair of entities were modeled with the objective of finding the optimal structural graph with at least 1 Markov lag until a maximum of 2 lags, since there is fewer data and since the higher peaks were around the lag 1 and 2, not considering the lag 0. Also to control the model complexity, because the more parents each node has the more the DBN structure is prone to overfitting, considering the amount of observed data we have and the two variables per area to model. This common problem known as curse of dimensionality, has a scientific principle saying the simplest model must be preferred, that is a problem-solving statement as know as Occam's razor principle. So the simplicity is assured by the parametrization of the DBN structure learning process defined in chapter 3, where each node could only get two parents from preceding timeslices maximum.

Since in the exploratory data analysis it was already noticeable that the toxic contamination events were extremely rare by looking at the percentages of high values of DSP toxins and phytoplankton above the limit on the more relevant entities, and even occurring more at the same time step on different production areas than within a shifted time-window period of 1 to 2 weeks, as it can be observed in the cross-correlation graphs where the peaks are all almost in the lag 0, the conditional probabilities are expected to be low for each possible variable state where the DSP toxin or phytoplankton concentration are high, and only a few times higher than the lowest state on very specific conditions of its neighbors states.

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. The main focus was to confirm the relations that the cross-correlation coefficients already pointed out to be high on 1 or 2 lagged time series, for example the phytoplankton in LAG and the phytoplankton in OLH2 in fig. 4.4, and 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. And if the correlation coefficient of this relation was high as it shows to be in the LAG-OLH2 time series cross-correlation, it means that the variable at lag -1 explains the other time series growth, i.e. the phytoplankton at OLH2 has a predictive linkage on the phytoplankton at LAG in the previous week, thus the expected dependency being like $dsp\_phyto\_Vale\_da\_Lama[0] \longrightarrow dsp\_phyto\_Fortaleza[1]$, 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 could 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.

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 $dsp\_phyto\_Fortaleza[1]$ 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, for example OLH2 to LAG, meaning that other areas and environmental factors may help on the proliferation and contamination as well. Also if the correlation analysis, showed considerably high positive coefficients on both directions, for example by looking at the DPCCA coefficient of toxins relations of Quatro Águas shifted 1 week backward (lag -1) from Fortaleza time series approximately 0.4, and the coefficient of Quatro Águas shifted 1 week backward (lag +1) approximately 0.5, and the DBN analysis also confirms that are events on both directions, there is not sufficient information to state that one area impacts more than the other with statistical assurance.

4.2.1 Vale da Lama (LAG) - Quatro Águas (TAV)

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, see fig. 4.7.

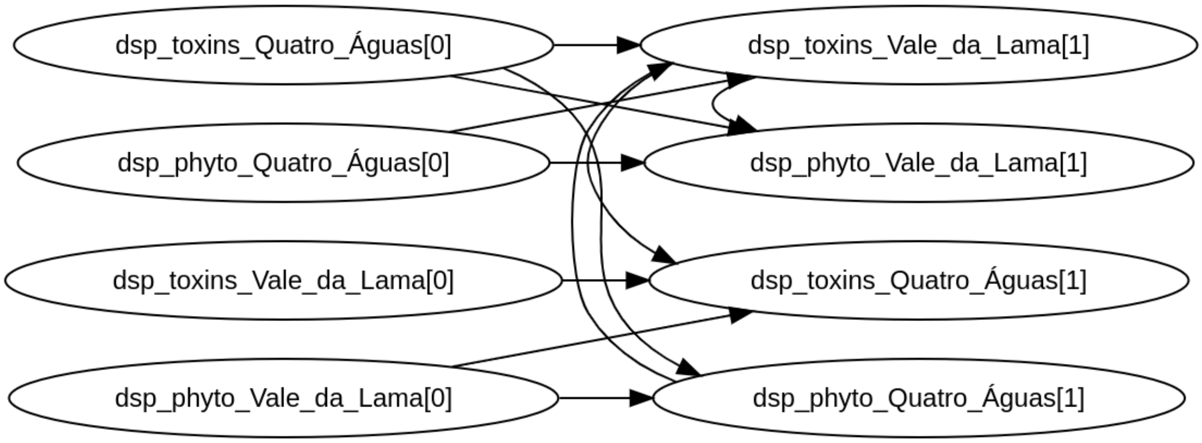
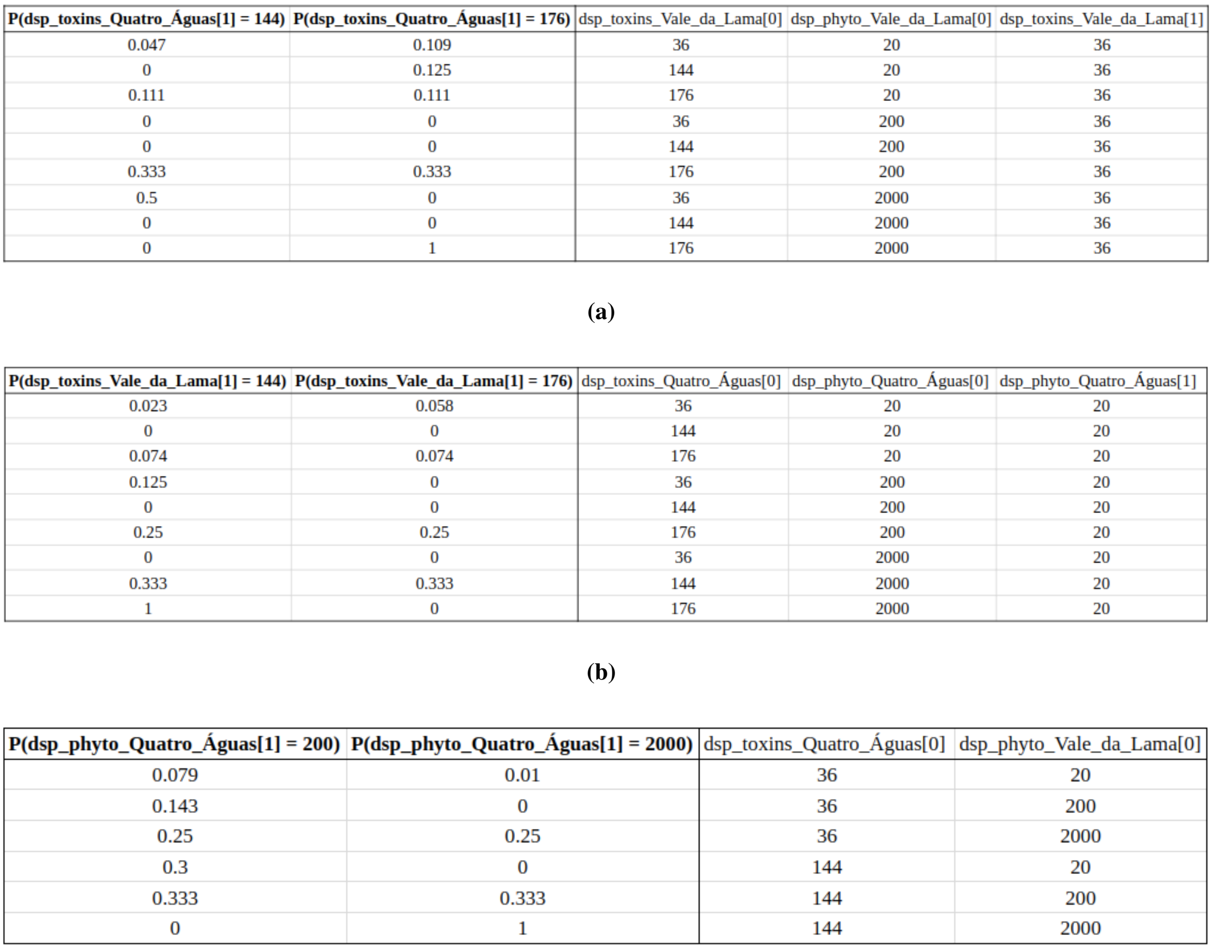

dsp_toxins_Quatro_Águas[0]
dsp_phyto_Quatro_Águas[0]
dsp_toxins_Vale_da_Lama[0]
dsp_phyto_Vale_da_Lama[0]
dsp_toxins_Vale_da_Lama[1]
dsp_phyto_Vale_da_Lama[1]
dsp_toxins_Quatro_Águas[1]
dsp_phyto_Quatro_Águas[1]

Figure 4.7: LAG-TAV DBN

By analyzing the *dsp_toxins_Vale_da_Lama[1]* and the *dsp_toxins_Quatro_Águas[1]* node sub-CPTs of the CPTs in appendix A, the sub-CPT in fig. 4.8 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).



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 4.8: 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

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.

4.2.2 Vale da Lama (LAG) - Fortaleza (OLH2)

For the pair LAG-OLH2 the inter-slice dependencies more expected were from *dsp_phyto_Vale_da_Lama[0]* to *dsp_phyto_Fortaleza[1]*, and from *dsp_toxins_Fortaleza[0]* to *dsp_toxins_Vale_da_Lama[1]*, and the DBN structure in fig. 4.9 confirms that dependencies.

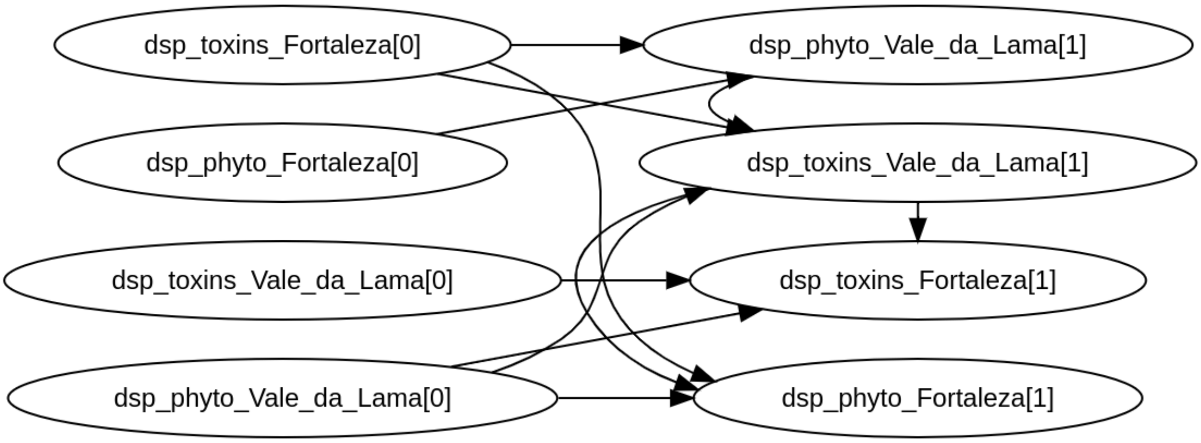
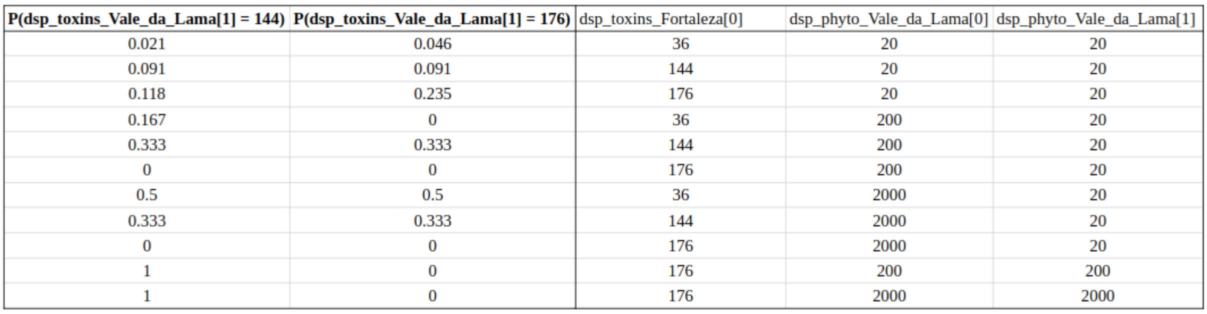

dsp_toxins_Fortaleza[0]
dsp_phyto_Fortaleza[0]
dsp_toxins_Vale_da_Lama[0]
dsp_phyto_Vale_da_Lama[0]
dsp_phyto_Vale_da_Lama[1]
dsp_toxins_Vale_da_Lama[1]
dsp_toxins_Fortaleza[1]
dsp_phyto_Fortaleza[1]

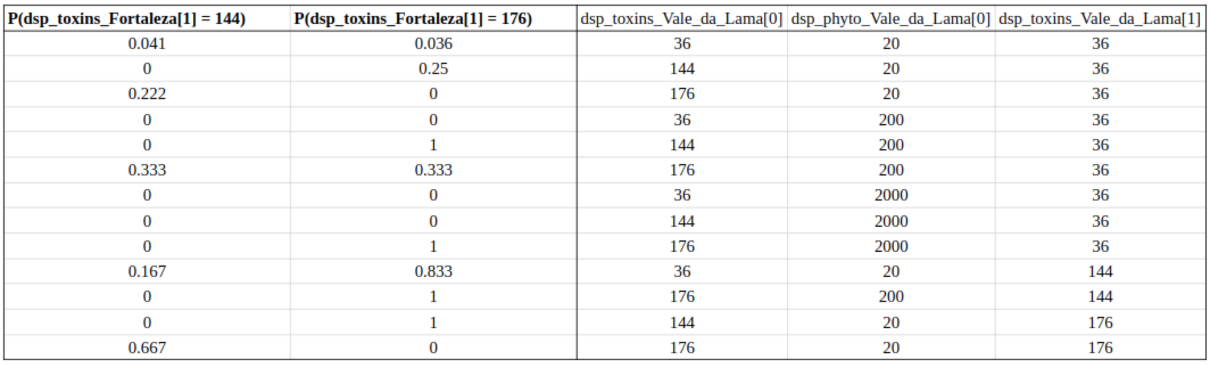
Figure 4.9: LAG-OLH2 DBN

First by analyzing the *dsp_toxins_Vale_da_Lama[1]* node sub-CPT of the CPT, there is a clear increase of the probability on both intervals of high toxin concentration at LAG as the toxins concentration at OLH2 rises, but it starts to be dependent of its own concentration of phytoplankton which could be normal since there should exist some DSP phyto concentration on the area itself. The other toxin relation in the opposite direction shows that there is a clear evidence that the DSP toxin concentration at OLH2 not only depends on the DSP toxin concentration at LAG, but also on the increase of the phytoplankton concentration at LAG, but now the probability of the higher value toxin concentration state, reaches 1 in two situations, without depending on the intra-slice variable, by nullifying its variation at state 36. But there is also evidence of occurring at the same time without depending only on the increase of the variables from preceding timeslices, with some specific conditions as it is shown in fig. 4.10(b). The dependency between *dsp_phyto_Vale_da_Lama[0]* and *dsp_phyto_Fortaleza[1]* that has a 0.7 correlation coefficient only has a case where the conditional probability of the phytoplankton concentration state 2000 is higher than the probability of the state 20, that is when the *dsp_phyto_Vale_da_Lama[0]* is equal to 2000.



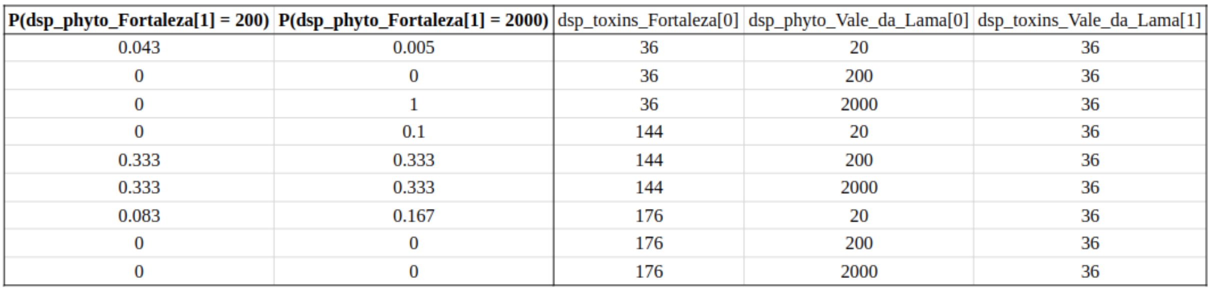
P(dsp_toxins_Vale_da_Lama[1] = 144)	P(dsp_toxins_Vale_da_Lama[1] = 176)	dsp_toxins_Fortaleza[0]	dsp_phyto_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[1]
0.021	0.046	36	20	20
0.091	0.091	144	20	20
0.118	0.235	176	20	20
0.167	0	36	200	20
0.333	0.333	144	200	20
0	0	176	200	20
0.5	0.5	36	2000	20
0.333	0.333	144	2000	20
0	0	176	2000	20
1	0	176	200	200
1	0	176	2000	2000

(a)



P(dsp_toxins_Fortaleza[1] = 144)	P(dsp_toxins_Fortaleza[1] = 176)	dsp_toxins_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]
0.041	0.036	36	20	36
0	0.25	144	20	36
0.222	0	176	20	36
0	0	36	200	36
0	1	144	200	36
0.333	0.333	176	200	36
0	0	36	2000	36
0	0	144	2000	36
0	1	176	2000	36
0.167	0.833	36	20	144
0	1	176	200	144
0	1	144	20	176
0.667	0	176	20	176

(b)



P(dsp_phyto_Fortaleza[1] = 200)	P(dsp_phyto_Fortaleza[1] = 2000)	dsp_toxins_Fortaleza[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]
0.043	0.005	36	20	36
0	0	36	200	36
0	1	36	2000	36
0	0.1	144	20	36
0.333	0.333	144	200	36
0.333	0.333	144	2000	36
0.083	0.167	176	20	36
0	0	176	200	36
0	0	176	2000	36

(c)

Figure 4.10: Two Sub-CPTs of DSP toxins at Vale da Lama (a) and at Fortaleza (b), and the Sub-CPT of DSP phyto at Fortaleza (c) in LAG-OLH2 DBN

4.2.3 Vale da Lama (LAG) - Monte Gordo (L9)

According to the cross-correlation analysis both from *dsp_toxins_Monte Gordo[0]* to *dsp_toxins_Vale_da_Lama[1]* and from *dsp_toxins_Vale_da_Lama[0]* to *dsp_toxins_Monte Gordo[1]*, and the positive peak at lag -1 on the relation from *dsp_phyto_Vale_da_Lama[0]* to *dsp_phyto_Monte Gordo[1]*, and the DBN structure analysis is focused on finding this impacts, and try to find other variables that may help explain the positive correlations.

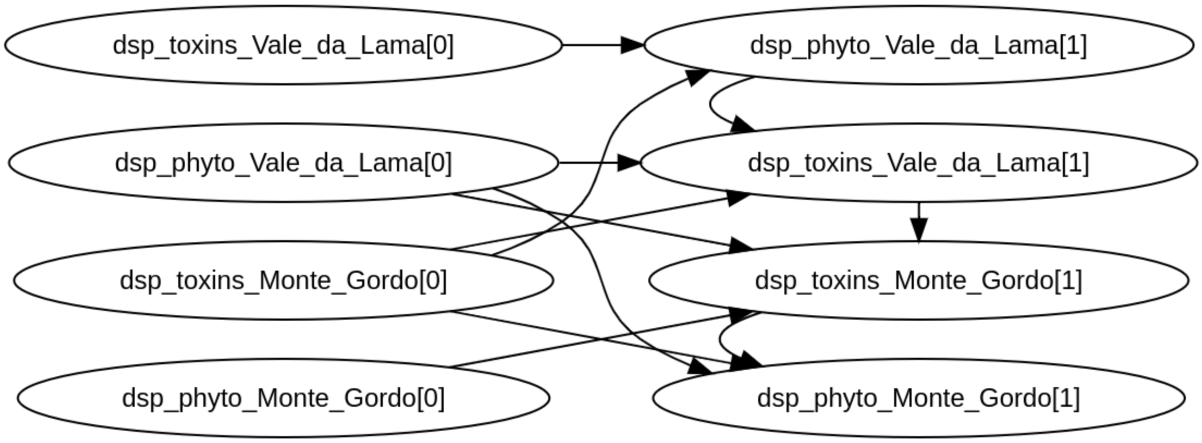

dsp_toxins_Vale_da_Lama[0]
dsp_phyto_Vale_da_Lama[1]
dsp_phyto_Vale_da_Lama[0]
dsp_toxins_Vale_da_Lama[1]
dsp_toxins_Monte_Gordo[0]
dsp_toxins_Monte_Gordo[1]
dsp_phyto_Monte_Gordo[0]
dsp_phyto_Monte_Gordo[1]

Figure 4.11: LAG-L9 DBN

First and foremost by looking at the CPTs in appendix A, there is clear evidence that the probability of DSP phytoplankton at Monte Gordo being at the higher states (200 or 2000), are equal to one third (0.333) only when the concentration of phytoplankton at Vale da Lama are at the highest concentration states (200 and 2000), apart from not existing a set of states conditions from its parents nodes with a probability higher for the higher concentration levels than the lowest level (20). The opposite direction of phytoplankton relation is not informative enough to be modeled as a dependence in the DBN in fig. 4.11, which may suggest that the expected impact direction happens more frequently indeed and with a positive correlation verified. The low correlation between toxins at L9 a week before and LAG in the next week, which was the highest coefficient from both directions, is confirmed by the lack of explainability of the DBN model when increasing only the variable *dsp_toxins_Monte_Gordo[0]* in the week before. And the two specific conditions where the toxin concentration in LAG is high with a conditional probability equal to 1, assuring that this type of contamination levels are only expressed in LAG when there is also a higher concentration of phytoplankton in LAG in the same timeslice. Also in the opposite direction, there is not any dependence on the preceding timeslice, i.e. from *dsp_toxins_Vale_da_Lama[0]* to *dsp_toxins_Monte Gordo[1]*.

4.2.4 Fortaleza (OLH2) - Quatro Águas (TAV)

Assuming that for both toxins and phyto time series, the correlations of lag +1 and lag -1, which is like saying the correlations from *dsp_toxins_Quatro_Águas[0]* to *dsp_toxins_Fortaleza[1]* and from *dsp_toxins_Fortaleza[0]* to *dsp_toxins_Quatro_Águas[1]*, and the same for the phytoplankton, were more or less equally positive for both influence directions and on both cross-correlation methods. So the dependency analysis was done to all DBN nodes.

Looking at the toxin concentration at Fortaleza CPT in fig. 4.13(a), there is enough evidence to show that

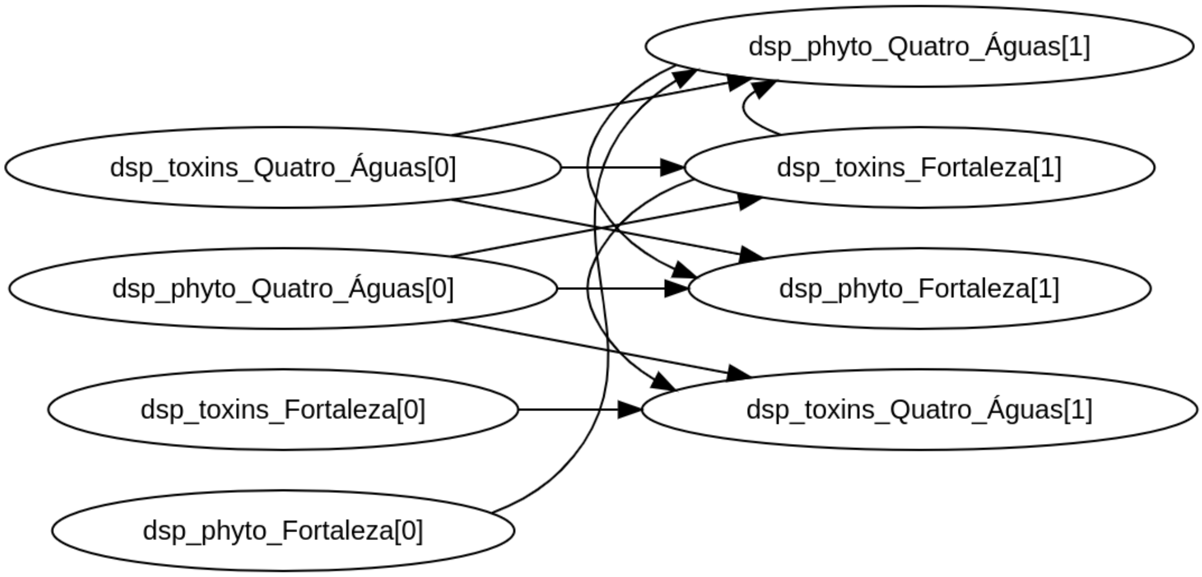
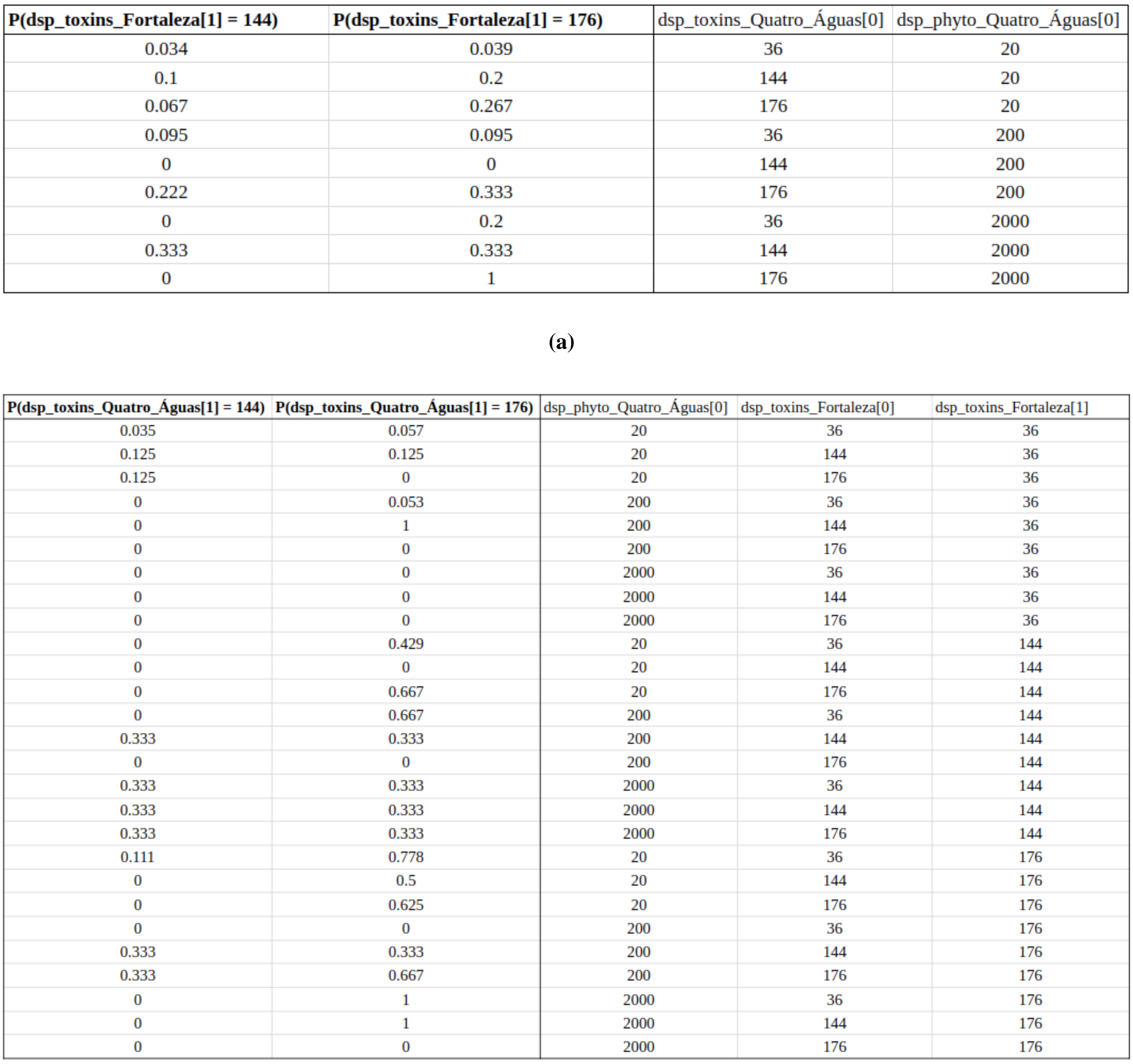

dsp_phyto_Quatro_Águas[1]
dsp_toxins_Quatro_Águas[0]
dsp_toxins_Fortaleza[1]
dsp_phyto_Quatro_Águas[0]
dsp_phyto_Fortaleza[1]
dsp_toxins_Fortaleza[0]
dsp_toxins_Quatro_Águas[1]
dsp_phyto_Fortaleza[0]

Figure 4.12: OLH2-TAV DBN

is highly probable to have high concentrations levels with a proportional increase of toxins in both toxins and phytoplankton at Quatro Águas, suggesting a lagged predictive link from TAV to OLH2 direction. The toxin concentration at Quatro Águas, has a dependence with a variable from the same timeslice, and the CPT in fig. 4.13(b) it shows more evidence on the toxin concentration state of 176 μg/kg at Quatro Águas. But by locking the variable in the same timeslice, only on the sub-CPT with toxins at Fortaleza on state 36, it is observed a proportional growth of the probability in state 176 at Quatro Águas when the toxins at Fortaleza in the week before rises. And unlike the opposite toxin relation, that only has parents from the preceding timeslice and only from its neighbor, the toxins in the next week at Quatro Águas still depend on its own state information, meaning Fortaleza has low information on toxins at Quatro Águas as expected has it had the lower correlation comparing with both lags -1 and +1. I've also analyzed both directions for phytoplankton concentration impact on both areas, and as it was expected both probabilities of high phytoplankton concentration levels (200 or 2000) increase proportionally with the phytoplankton concentration at the respective neighbor in the sub-CPTs in fig. 4.14(a) and fig. 4.14(b), and both have some cases of the probability being higher on the higher states than on the lowest state (20).



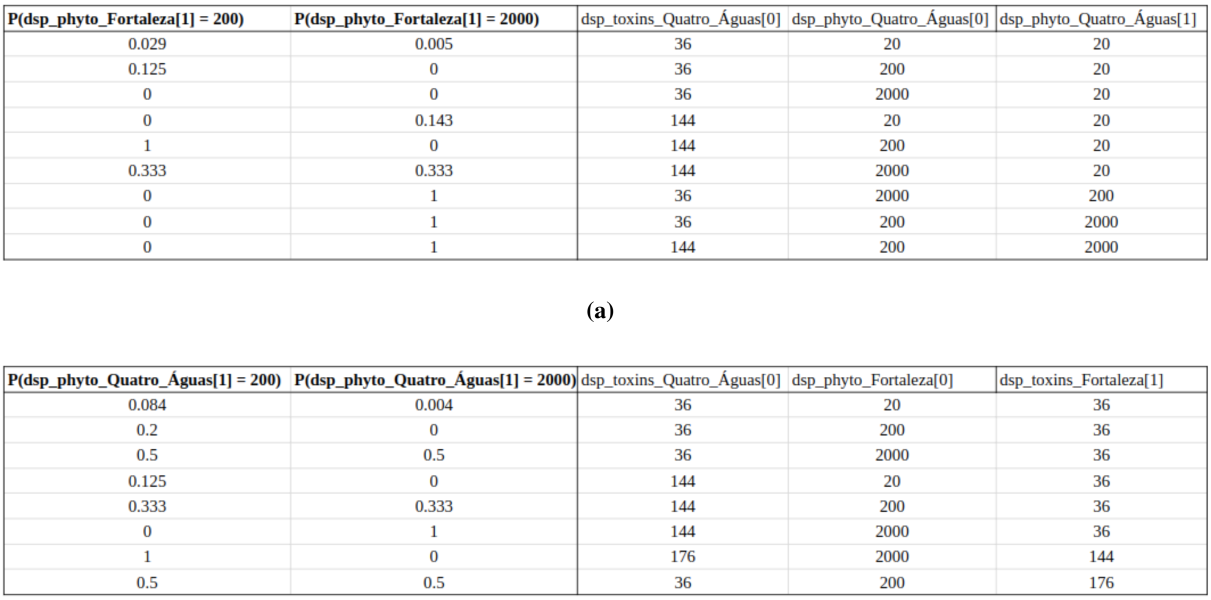
P(dsp_toxins_Fortaleza[1] = 144)	P(dsp_toxins_Fortaleza[1] = 176)	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]
0.034	0.039	36	20
0.1	0.2	144	20
0.067	0.267	176	20
0.095	0.095	36	200
0	0	144	200
0.222	0.333	176	200
0	0.2	36	2000
0.333	0.333	144	2000
0	1	176	2000

(a)

P(dsp_toxins_Quatro_Águas[1] = 144)	P(dsp_toxins_Quatro_Águas[1] = 176)	dsp_phyto_Quatro_Águas[0]	dsp_toxins_Fortaleza[0]	dsp_toxins_Fortaleza[1]
0.035	0.057	20	36	36
0.125	0.125	20	144	36
0.125	0	20	176	36
0	0.053	200	36	36
0	1	200	144	36
0	0	200	176	36
0	0	2000	36	36
0	0	2000	144	36
0	0	2000	176	36
0	0.429	20	36	144
0	0	20	144	144
0	0.667	20	176	144
0	0.667	200	36	144
0.333	0.333	200	144	144
0	0	200	176	144
0.333	0.333	2000	36	144
0.333	0.333	2000	144	144
0.333	0.333	2000	176	144
0.111	0.778	20	36	176
0	0.5	20	144	176
0	0.625	20	176	176
0	0	200	36	176
0.333	0.333	200	144	176
0.333	0.667	200	176	176
0	1	2000	36	176
0	1	2000	144	176
0	0	2000	176	176

(b)

Figure 4.13: Two Sub-CPTs of DSP toxins at Fortaleza (a) and at Quatro Águas (b) in OLH2-TAV DBN



P(dsp_phyto_Fortaleza[1] = 200)	P(dsp_phyto_Fortaleza[1] = 2000)	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]
0.029	0.005	36	20	20
0.125	0	36	200	20
0	0	36	2000	20
0	0.143	144	20	20
1	0	144	200	20
0.333	0.333	144	2000	20
0	1	36	2000	200
0	1	36	200	2000
0	1	144	200	2000

(a)

P(dsp_phyto_Quatro_Águas[1] = 200)	P(dsp_phyto_Quatro_Águas[1] = 2000)	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Fortaleza[0]	dsp_toxins_Fortaleza[1]
0.084	0.004	36	20	36
0.2	0	36	200	36
0.5	0.5	36	2000	36
0.125	0	144	20	36
0.333	0.333	144	200	36
0	1	144	2000	36
1	0	176	2000	144
0.5	0.5	36	200	176

(b)

Figure 4.14: Two sub-CPTs of DSP phyto at Fortaleza (a) and at Quatro Águas (b) in OLH2-TAV DBN

4.2.5 Fortaleza (OLH2) - Monte Gordo (L9)

Assuming the highest correlations, the expected inter-slice dependencies were from *dsp_phyto_Fortaleza[0]* to *dsp_phyto_Monte Gordo[1]*, and from *dsp_toxins_Monte Gordo[0]* to *dsp_toxins_Fortaleza[1]*, and the phytoplankton dependency is the only one the DBN structure does not learn.

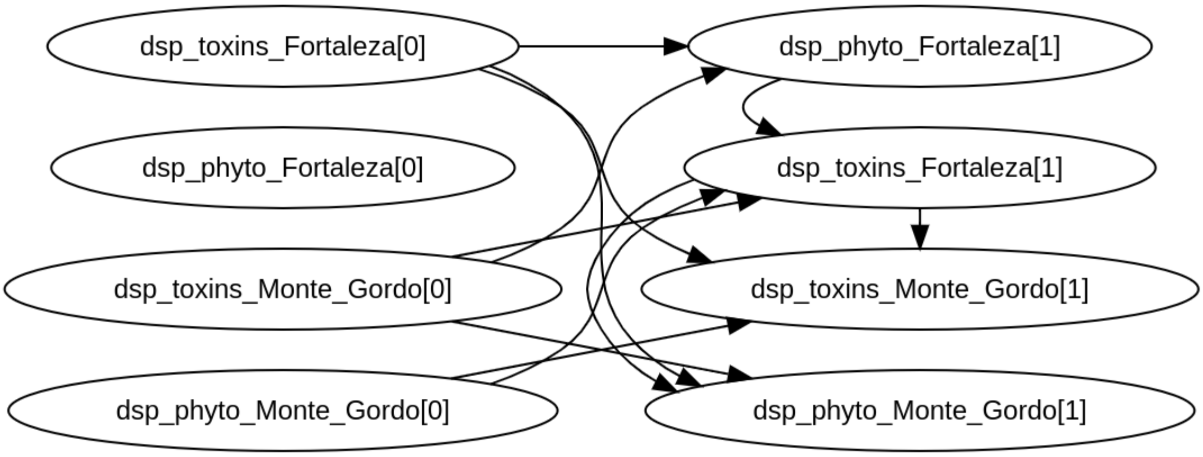
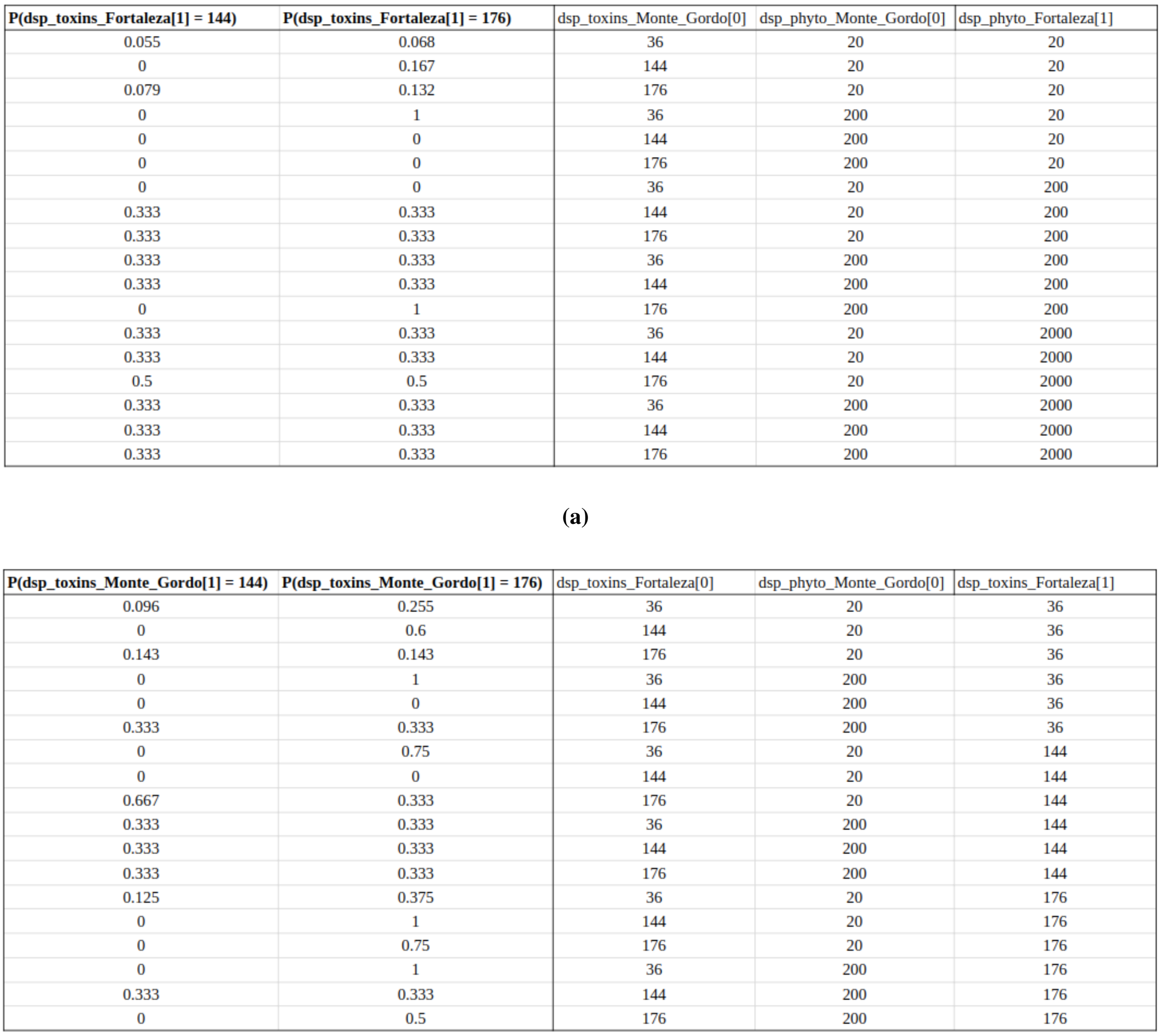


Figure 4.15: OLH2-L9 DBN

On the expected toxin impact direction with higher positive relation, in fig. 4.16(a), it seems to exist two set of conditions where the probability of the DSP toxins concentration are higher than the probability of the lower state, apart from the case when the probability equals to 1 on state 176 when the toxin concentration at Monte Gordo is at state 36. The probability of getting high toxin concentration in L9, in fig. 4.16(b), also gets multiple cases of existing high concentration of toxins at Monte Gordo when there was both low or high concentration of DSP toxins at Fortaleza, but the probability of toxin concentration at Monte Gordo on state 144 increases proportionally to the concentration in Fortaleza, needing also the increase of the second variable of the previous timeslice, going from 0.096, to 0.143, 0.333 and to 0.667, with the dependency of increasing the variable in the same timestep. These analysis suggests that there is no clear evidence of a noticeable impact on a certain direction.



P(dsp_toxins_Fortaleza[1] = 144)	P(dsp_toxins_Fortaleza[1] = 176)	dsp_toxins_Monte_Gordo[0]	dsp_phyto_Monte_Gordo[0]	dsp_phyto_Fortaleza[1]
0.055	0.068	36	20	20
0	0.167	144	20	20
0.079	0.132	176	20	20
0	1	36	200	20
0	0	144	200	20
0	0	176	200	20
0	0	36	20	200
0.333	0.333	144	20	200
0.333	0.333	176	20	200
0.333	0.333	36	200	200
0.333	0.333	144	200	200
0	1	176	200	200
0.333	0.333	36	20	2000
0.333	0.333	144	20	2000
0.5	0.5	176	20	2000
0.333	0.333	36	200	2000
0.333	0.333	144	200	2000
0.333	0.333	176	200	2000

(a)

P(dsp_toxins_Monte_Gordo[1] = 144)	P(dsp_toxins_Monte_Gordo[1] = 176)	dsp_toxins_Fortaleza[0]	dsp_phyto_Monte_Gordo[0]	dsp_toxins_Fortaleza[1]
0.096	0.255	36	20	36
0	0.6	144	20	36
0.143	0.143	176	20	36
0	1	36	200	36
0	0	144	200	36
0.333	0.333	176	200	36
0	0.75	36	20	144
0	0	144	20	144
0.667	0.333	176	20	144
0.333	0.333	36	200	144
0.333	0.333	144	200	144
0.333	0.333	176	200	144
0.125	0.375	36	20	176
0	1	144	20	176
0	0.75	176	20	176
0	1	36	200	176
0.333	0.333	144	200	176
0	0.5	176	200	176

(b)

Figure 4.16: Two sub-CPTs of DSP toxins at Fortaleza (a) and at Monte Gordo (b) in OLH2-L9 DBN

4.2.6 Quatro Águas (TAV) - Monte Gordo (L9)

The expected inter-slice dependencies were from *dsp_phyto_Quatro_Águas[0]* to *dsp_phyto_Monte Gordo[1]*, and from *dsp_toxins_Quatro_Águas[0]* to *dsp_toxins_Monte Gordo[1]* or from *dsp_toxins_Monte Gordo[0]* to *dsp_toxins_Quatro_Águas[1* since the correlations on lag -1 and +1 are more or less equally correlated on both cross-correlation methods.

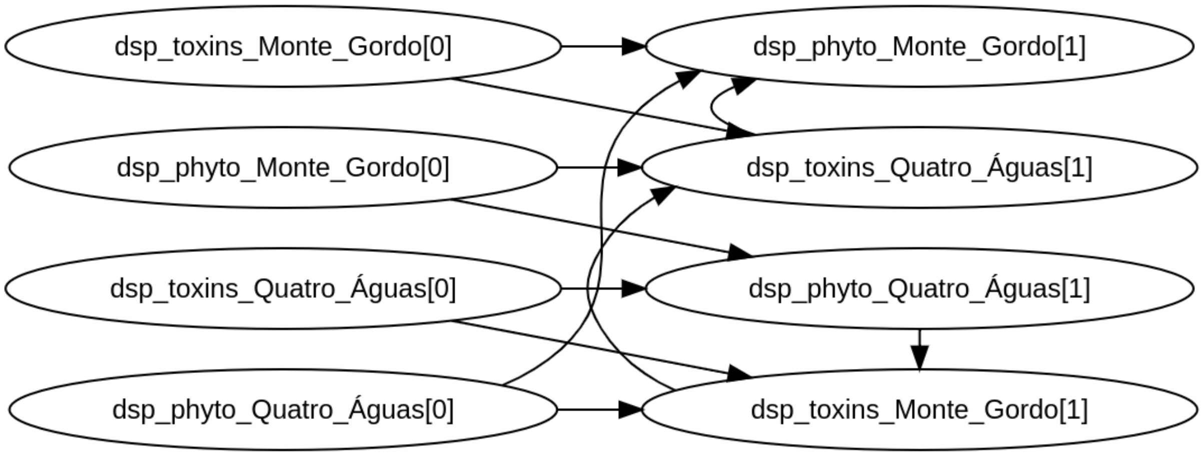
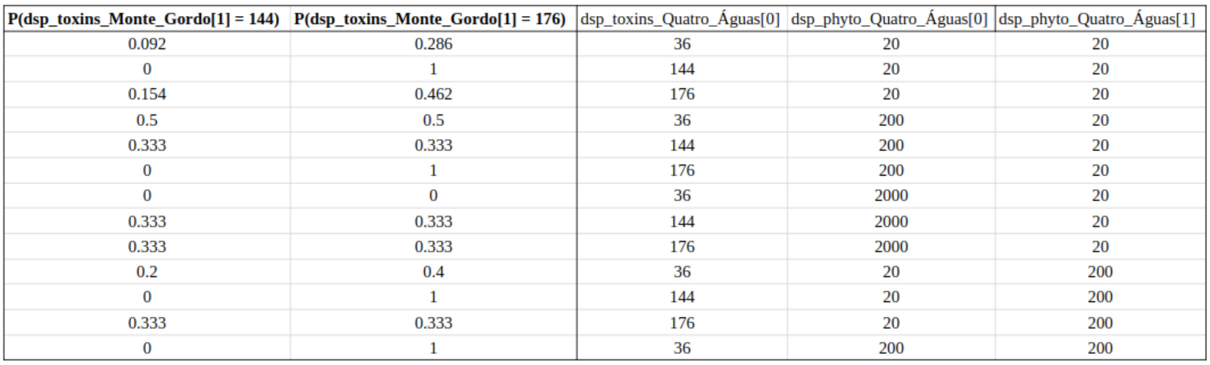

dsp_toxins_Monte_Gordo[0]
dsp_phyto_Monte_Gordo[1]
dsp_phyto_Monte_Gordo[0]
dsp_toxins_Quatro_Águas[1]
dsp_toxins_Quatro_Águas[0]
dsp_phyto_Quatro_Águas[1]
dsp_phyto_Quatro_Águas[0]
dsp_toxins_Monte_Gordo[1]

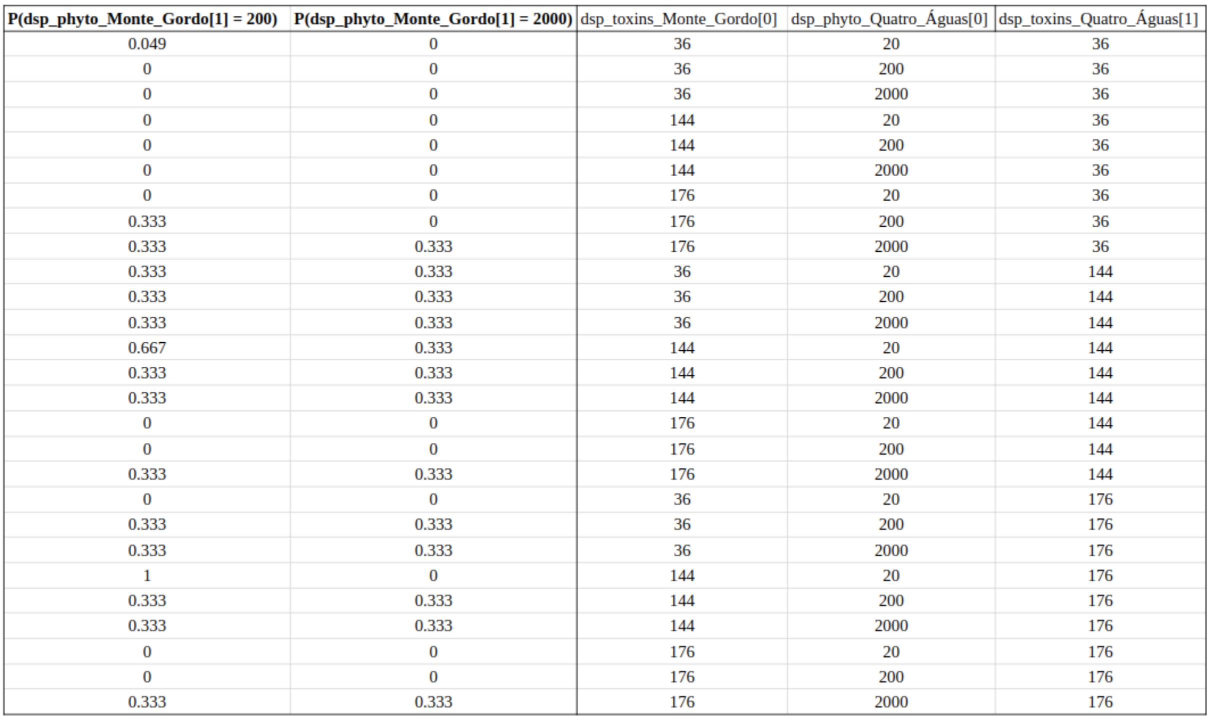
Figure 4.17: TAV-L9 DBN

For the expected toxins impact relations, both are only dependent on variables of the neighbor, and apart from both having some high conditional probabilities of higher concentrations when its neighbors are on the lowest concentration state, the probability of DSP toxins at Monte Gordo on state 176, shows some proportional growth from 0.286, to 0.462 and finally to 1 with toxins above the limit at Quatro Águas and the phytoplankton as well, as shown in fig. 4.18(a). The phytoplankton relation from TAV to L9 that had the highest correlation between lag -1 and lag +1 on both correlation methods, have a stochastic conditional probabilities distribution since for almost all the possible states the probability is 0.333 independently of its parents conditions, as observed in fig. 4.18(b), except for one set of conditions, where the state of phytoplaknton was at 20 at Quatro Águas, where the probability is higher than the other states.



P(dsp_toxins_Monte_Gordo[1] = 144)	**P(dsp_toxins_Monte_Gordo[1] = 176)**	dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]
0.092	0.286	36	20	20
0	1	144	20	20
0.154	0.462	176	20	20
0.5	0.5	36	200	20
0.333	0.333	144	200	20
0	1	176	200	20
0	0	36	2000	20
0.333	0.333	144	2000	20
0.333	0.333	176	2000	20
0.2	0.4	36	20	200
0	1	144	20	200
0.333	0.333	176	20	200
0	1	36	200	200

(a)



P(dsp_phyto_Monte_Gordo[1] = 200)	**P(dsp_phyto_Monte_Gordo[1] = 2000)**	dsp_toxins_Monte_Gordo[0]	dsp_phyto_Quatro_Águas[0]	dsp_toxins_Quatro_Águas[1]
0.049	0	36	20	36
0	0	36	200	36
0	0	36	2000	36
0	0	144	20	36
0	0	144	200	36
0	0	144	2000	36
0	0	176	20	36
0.333	0	176	200	36
0.333	0.333	176	2000	36
0.333	0.333	36	20	144
0.333	0.333	36	200	144
0.333	0.333	36	2000	144
0.667	0.333	144	20	144
0.333	0.333	144	200	144
0.333	0.333	144	2000	144
0	0	176	20	144
0	0	176	200	144
0.333	0.333	176	2000	144
0	0	36	20	176
0.333	0.333	36	200	176
0.333	0.333	36	2000	176
1	0	144	20	176
0.333	0.333	144	200	176
0.333	0.333	144	2000	176
0	0	176	20	176
0	0	176	200	176
0.333	0.333	176	2000	176

(b)

Figure 4.18: One sub-CPT of DSP toxins at Monte Gordo (a) and the sub-CPT of DSP phyto at Quatro Águas (b) in TAV-L9 DBN

Chapter 5

Conclusion and Future Work

The necessity of anticipating the shellfish contamination event can be tackled from different ways, i.e. multiple different solutions that complement each other in terms of modeling the phenomenon and get an theoretical explanation as complete as possible. This project comes as a complement for regression predictive models, where the level of the concentration is forecast from feeding an approximated function with multivariate time series. 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.

Apart from not being able to assume with high certainty the impact between the areas in the south coast, the expected sea current direction in terms of phytoplankton dispersion was more or less confirmed, and the advantage of creating this model is that it creates a compressed way of visualizing the conditional distributions and take insights from which are the concentration of toxins and phytoplankton on a neighboring area that may translate on a toxic event in the week after that. Which could be a predictive tool to estimate the probability of occurring biotoxin contamination on shellfish and being continuously updating the likelihoods to get a more precise prediction, by using the prediction feature of MAESTRO and training again the DBN model respectively. Also it can be used 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.

To get more interesting inter-dependencies results, it can be added the already pre-processed and integrated meteorological time series of IPMA, the oceanographic ones from MARETEC, and the imagery sensor variables already extracted from Copericus products, which can be updated to have the latest data. And also the DBN training process can have different configuration, as the markov lag being increased to 2 lags, studying a more sparse effect throughout time, since other variables can be added to the network. As future work, the pipeline of the data processing step is complete and available at my GitHub page, which I can give access by request, where all the exploratory data analysis and data preparation for MAESTRO format can be replicated and altered on the last modified Jupyter notebook. A possible future hypothesis to be verified is knowing that if the wind and sea current direction predominance coming from the North would potentiate phytoplankton blooms dispersion from Caparica or Trafaria to Jangada, there were high conditional probabilities of having high concentration of toxins passing two weeks on Jangada from the high concentration on Caparica or Trafaria. From the sea surface temperature I would expect a intra-slice dependence between it and phytoplankton or maybe between that same phyto variable but 1 week after getting high temperatures. If there isn't that kind of results the justification could be that the oceonographic behavior between coastal areas and lagoons is more complex, and this must be taking into account, but also the lack of data and all the methods applied to data pre-processing, such as imputation, discretization can influence the results. Furthermore 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. Both products are sources of data that can be gathered on a more adequate way, to model the phytoplankton dispersion and other environmental factors influence through other intermediary locations, mainly from more offshore data. This intermediate points between potential pairs of areas could help to model the actual direction of phytoplankton dispersion with more precision.

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.

Bibliography

[1] J. Guillen, F. Natale, N. Carvalho, J. Casey, J. Hofherr, J.-N. Druon, G. Fiore, M. Gibin, A. Zanzi, and J. T. Martinsohn, “Global seafood consumption footprint,” Ambio, vol. 48, p. 111–122, 2019.

[2] P. Guillotreau, V. L. Bihan, B. Morineau, and S. Pardo, “The vulnerability of shellfish farmers to HAB events: An optimal matching analysis of closure decrees,” Harmful Algae, vol. 101, p. 101968, 01 2021, [CrossRef].

[3] C. J. Gobler, “Climate change and harmful algal blooms: Insights and perspective,” Harmful Algae, vol. 91, p. 101731, 01 2020, climate change and harmful algal blooms.

[4] Griffith, “Ocean warming along temperate western boundaries of the northern hemisphere promotes an expansion of cochlodinium polykrikoides blooms,” Proceedings of the Royal Society B: Biological Sciences, vol. 286, no. 1904, p. 20190340, 06 2019, [CrossRef].

[5] K. H. Lee, H. J. Jeong, K. Lee, P. J. Franks, K. A. Seong, S. Y. Lee, M. J. Lee, S. Hyeon Jang, E. Potvin, A. Suk Lim, E. Y. Yoon, Y. D. Yoo, N. S. Kang, and K. Y. Kim, “Effects of warming and eutrophication on coastal phytoplankton production,” Harmful Algae, vol. 81, pp. 106–118, 01 2019, [CrossRef].

[6] R. C. Cruz, P. Reis Costa, S. Vinga, L. Krippahl, and M. B. Lopes, “A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination,” Journal of Marine Science and Engineering, vol. 9, no. 3, 2021, [CrossRef].

[7] N. F. L. Kenneth R. Hinga, Heeseon Jeon, “Marine eutrophication review,” 1995, u.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, Coastal Ocean Office. [Online]. Available: https://repository.library.noaa.gov/view/noaa/2911

[8] 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, vol. 116, pp. 231–240, 06 2017, [CrossRef].

[9] D. M. Anderson, P. M. Glibert, and J. M. Burkholder, “Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences,” Estuaries and Coasts, vol. 25, no. 4, pp. 704–726, 2002, [CrossRef].

[10] T. Trombetta, F. Vidussi, S. Mas, D. Parin, M. Simier, and B. Mostajir, “Water temperature drives phytoplankton blooms in coastal waters,” PLOS ONE, vol. 14, no. 4, pp. 1–28, 04 2019, [CrossRef].

[11] T. T. V. Tong, T. H. H. Le, B. M. Tu, and D. C. Le, “Spatial and seasonal variation of diarrheic shellfish poisoning (dsp) toxins in bivalve mollusks from some coastal regions of vietnam and assessment of potential health risks,” Marine Pollution Bulletin, vol. 133, pp. 911–919, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0025326X1830448X

[12] 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, vol. 7, no. 1, pp. 11–25, 2008, [CrossRef].

[13] E. Ansa, H. Lubberding, J. Ampofo, and H. Gijzen, “The role of algae in the removal of escherichia coli in a tropical eutrophic lake,” Ecological Engineering, vol. 37, no. 2, pp. 317–324, 2011, [CrossRef].

[14] R. A. Costa, “Escherichia coli in seafood: A brief overview.” Advances in Bioscience and Biotechnology, vol. 4, pp. 450–454, 03 2013, [CrossRef].

[15] L. M. Grattan, S. Holobaugh, and J. G. Morris, “Harmful algal blooms and public health,” Harmful Algae, vol. 57, pp. 2–8, 2016, [CrossRef].

[16] 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, applied Simulations and Integrated Modelling for the Understanding of Toxic and Harmful Algal Blooms (ASIMUTH).

[17] European Commission. Commission Regulation (EC) No 853/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific hygiene rules for on the hygiene of foodstuffs. Off. J. Eur. Union L 2004, 139, 55–205.

[18] European Commission. Commission Regulation (EC) No 854/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific rules for the organisation of official controls on products of animal origin intended for human consumption. Off. J. Eur. Union L 2004, 139, 206–320.

[19] J. A. Fernandes-Salvador, K. Davidson, M. Sourisseau, M. Revilla, W. Schmidt, D. Clarke, P. I. Miller, P. Arce, R. Fernández, L. Maman, A. Silva, C. Whyte, M. Mateo, P. Neira, M. Mateus, M. Ruiz-Villarreal, L. Ferrer, and J. Silke, “Current status of forecasting toxic harmful algae for the north-east atlantic shellfish aquaculture industry,” Frontiers in Marine Science, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2021.666583

[20] Https://www.copernicus.eu/en - European Union’s Earth Observation Programme.

[21] M. Mateus, A. D. Silva, H. de Pablo, M. Moita, T. Quental, and L. Pinto, Using Lagrangian Elements to simulate alongshore transport of Harmful Algal Blooms, 10 2013, pp. 235–248, [CrossRef].

[22] 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, vol. 112, p. 102171, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568988321002018

[23] 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, vol. 10, 2018. [Online]. Available: https://www.int-res.com/abstracts/aei/v10/p173-185/

[24] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications (Springer Texts in Statistics). Berlin, Heidelberg: Springer-Verlag, 2005, [CrossRef].

[25] 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, vol. 5, no. 1, pp. 1–7, 2015.

[26] D. Koller and N. Friedman, Probabilistic graphical models: principles and techniques, ser. Adaptive computation and machine learning. MIT Press, 2009.

[27] "Learning Bayesian networks: The combination of knowledge and statistical data," Machine Learning, vol. 20, pp. 197–243, 1995, [CrossRef].

[28] K. Shan, M. Shang, B. Zhou, L. Li, X. Wang, H. Yang, and L. Song, "Application of Bayesian network including microcystis morphospecies for microcystin risk assessment in three cyanobacterial bloom-plagued lakes, china," Harmful Algae, vol. 83, pp. 14–24, 2019, [CrossRef].

[29] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," IEEE Transactions on Information Theory, vol. 14, no. 3, pp. 462–467, 1968.

[30] 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 Science & Technology, vol. 46, no. 13, pp. 7283–7292, 2012, [CrossRef].

[31] 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, vol. 84, pp. 89–103, 2016, [CrossRef].

[32] J. L. Serras, S. Vinga, and A. M. Carvalho, "Outlier detection for multivariate time series using dynamic Bayesian networks," Applied Sciences, vol. 11, no. 4, 2021, [CrossRef].

[33] A. Alkawri, "Seasonal variation in composition and abundance of harmful dinoflagellates in yemeni waters, southern red sea," Marine Pollution Bulletin, vol. 112, no. 1, pp. 225–234, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0025326X16306488

[34] R. Salas and D. Clarke, "Review of dsp toxicity in ireland: Long-term trend impacts, biodiversity and toxin profiles from a monitoring perspective," Toxins, vol. 11, no. 2, 2019. [Online]. Available: https://www.mdpi.com/2072-6651/11/2/61

[35] 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, vol. 41, no. 2, pp. 187–197, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041010102002763

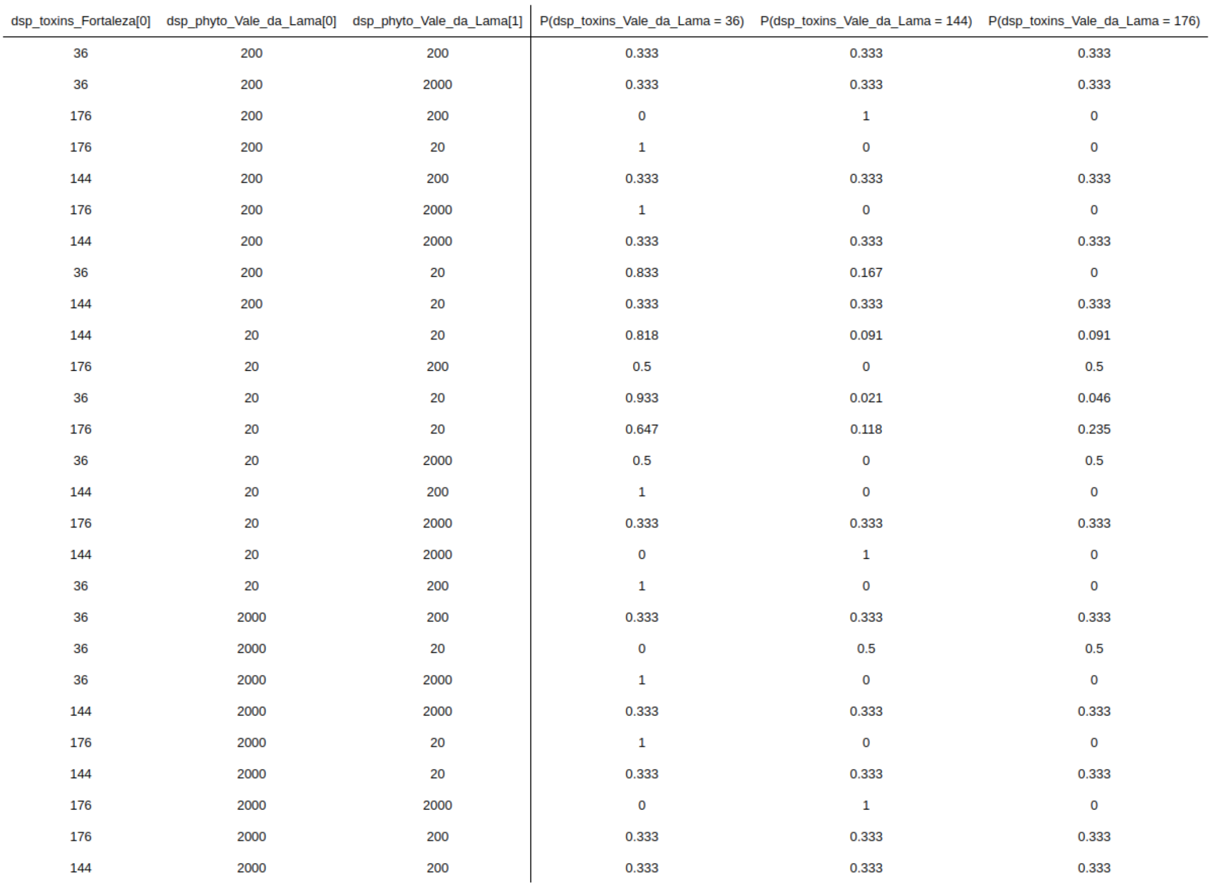
[36] A. A. Chadegani and D. Poursina, "An examination of the effect of discretization on a naive bayes models performance," Scientific research and essays, vol. 8, no. 44, pp. 2181–2186, 2013.

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

Appendix A

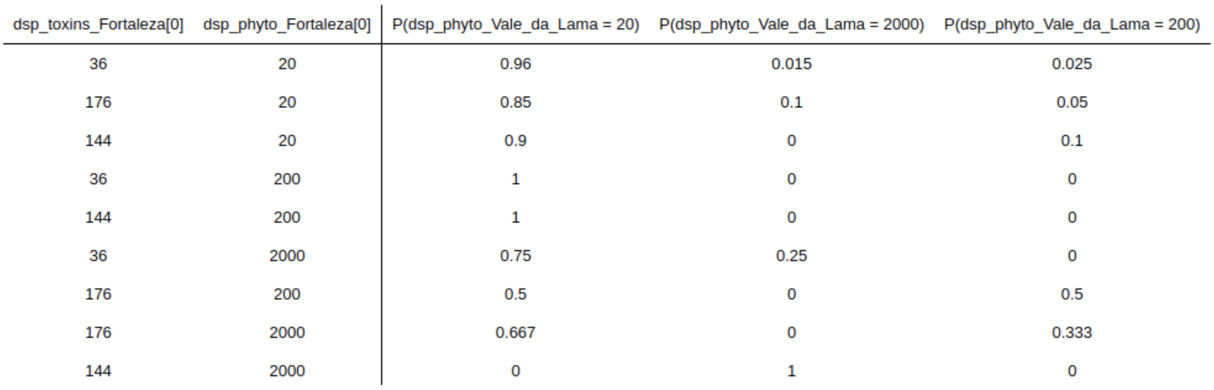
Complete Conditional Probabilities Tables

A.1 LAG-OLH2 DBN



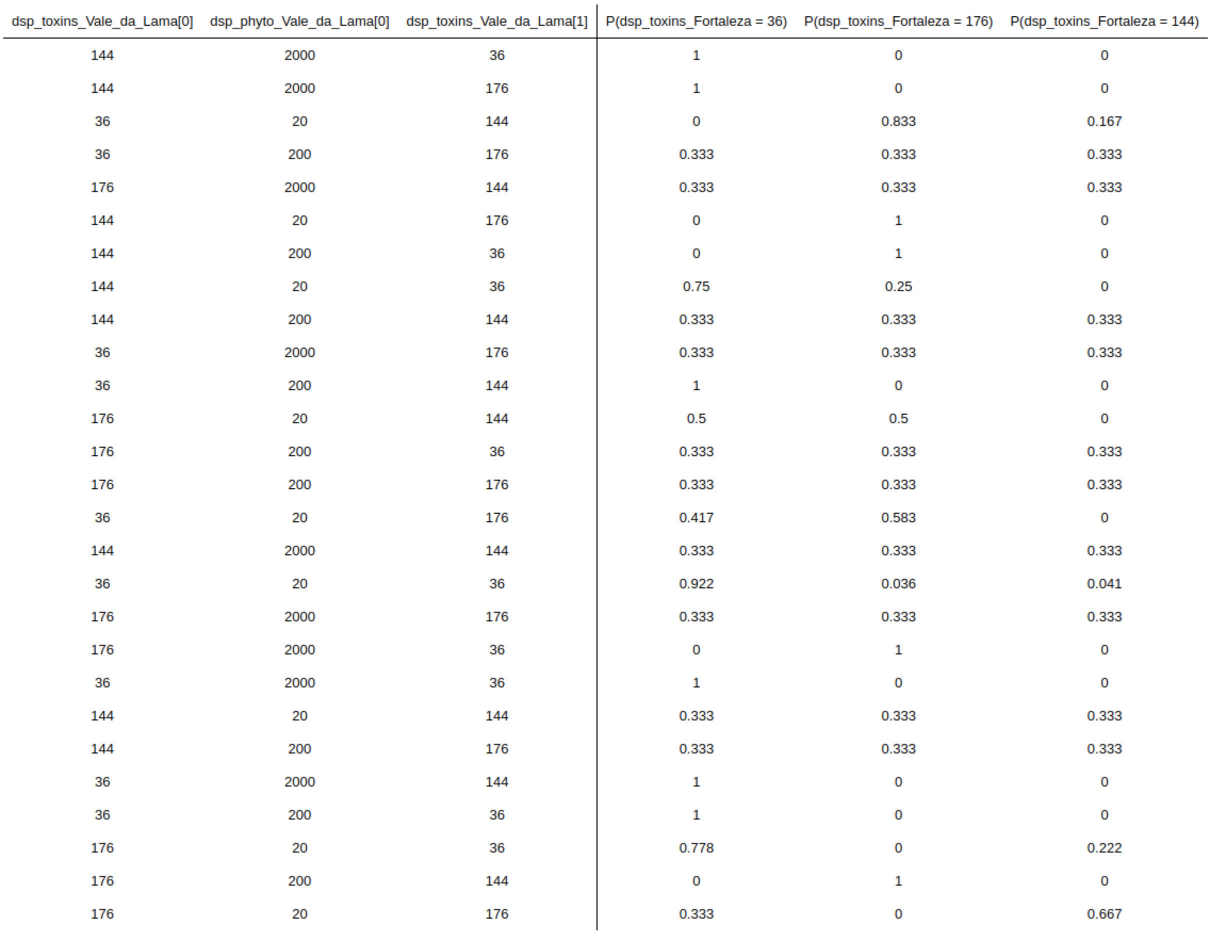
dsp_toxins_Fortaleza[0]	dsp_phyto_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[1]	P(dsp_toxins_Vale_da_Lama = 36)	P(dsp_toxins_Vale_da_Lama = 144)	P(dsp_toxins_Vale_da_Lama = 176)
36	200	200	0.333	0.333	0.333
36	200	2000	0.333	0.333	0.333
176	200	200	0	1	0
176	200	20	1	0	0
144	200	200	0.333	0.333	0.333
176	200	2000	1	0	0
144	200	2000	0.333	0.333	0.333
36	200	20	0.833	0.167	0
144	200	20	0.333	0.333	0.333
144	20	20	0.818	0.091	0.091
176	20	200	0.5	0	0.5
36	20	20	0.933	0.021	0.046
176	20	20	0.647	0.118	0.235
36	20	2000	0.5	0	0.5
144	20	200	1	0	0
176	20	2000	0.333	0.333	0.333
144	20	2000	0	1	0
36	20	200	1	0	0
36	2000	200	0.333	0.333	0.333
36	2000	20	0	0.5	0.5
36	2000	2000	1	0	0
144	2000	2000	0.333	0.333	0.333
176	2000	20	1	0	0
144	2000	20	0.333	0.333	0.333
176	2000	2000	0	1	0
176	2000	200	0.333	0.333	0.333
144	2000	200	0.333	0.333	0.333

Figure A.1: CPT of DSP Toxins Concentration at LAG conditioned in timeslice 1, with pair OLH2



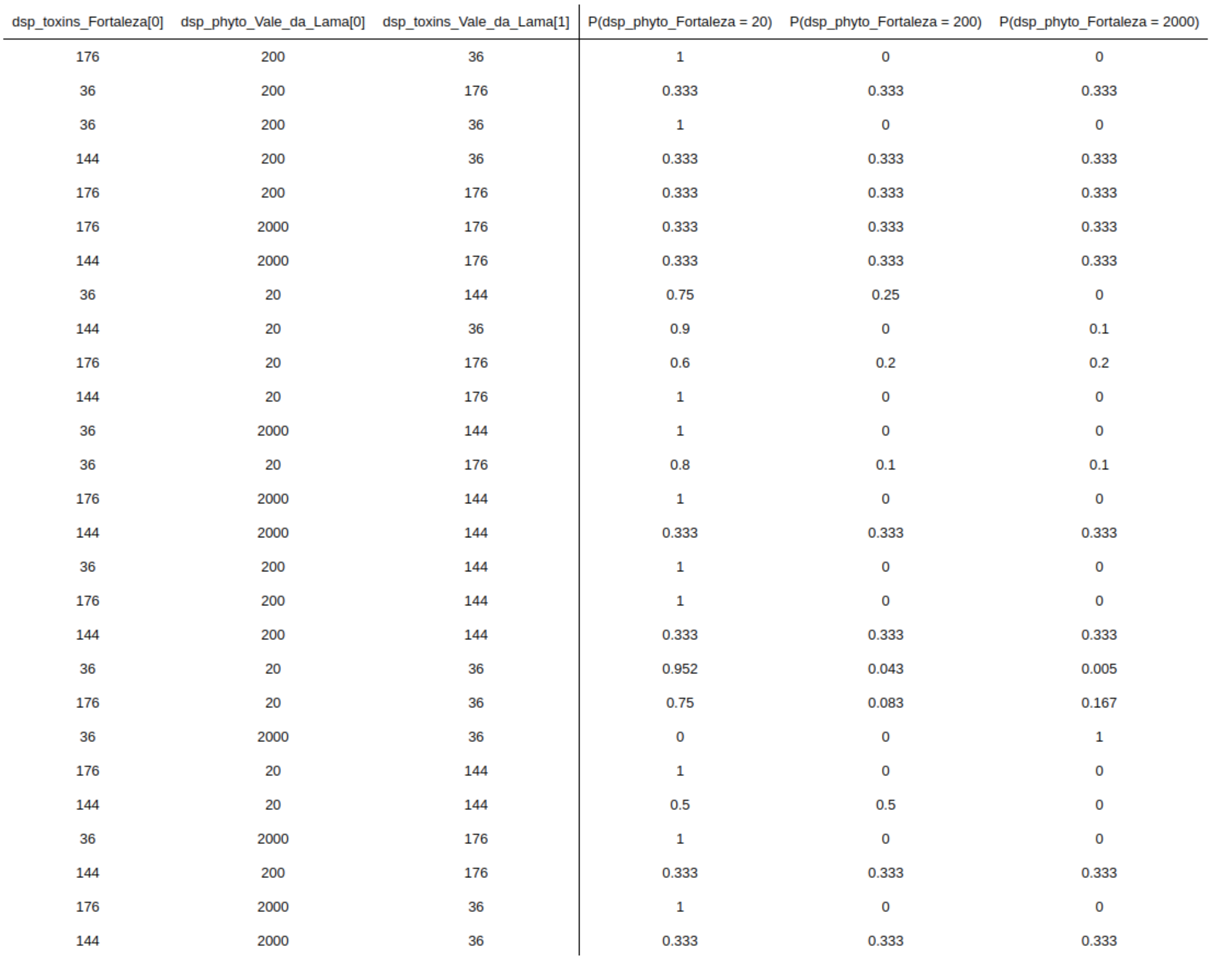
dsp_toxins_Fortaleza[0]	dsp_phyto_Fortaleza[0]	P(dsp_phyto_Vale_da_Lama = 20)	P(dsp_phyto_Vale_da_Lama = 2000)	P(dsp_phyto_Vale_da_Lama = 200)
36	20	0.96	0.015	0.025
176	20	0.85	0.1	0.05
144	20	0.9	0	0.1
36	200	1	0	0
144	200	1	0	0
36	2000	0.75	0.25	0
176	200	0.5	0	0.5
176	2000	0.667	0	0.333
144	2000	0	1	0

Figure A.2: CPT of DSP Phyto Concentration at LAG conditioned in timeslice 1, with pair OLH2



dsp_toxins_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]	P(dsp_toxins_Fortaleza = 36)	P(dsp_toxins_Fortaleza = 176)	P(dsp_toxins_Fortaleza = 144)
144	2000	36	1	0	0
144	2000	176	1	0	0
36	20	144	0	0.833	0.167
36	200	176	0.333	0.333	0.333
176	2000	144	0.333	0.333	0.333
144	20	176	0	1	0
144	200	36	0	1	0
144	20	36	0.75	0.25	0
144	200	144	0.333	0.333	0.333
36	2000	176	0.333	0.333	0.333
36	200	144	1	0	0
176	20	144	0.5	0.5	0
176	200	36	0.333	0.333	0.333
176	200	176	0.333	0.333	0.333
36	20	176	0.417	0.583	0
144	2000	144	0.333	0.333	0.333
36	20	36	0.922	0.036	0.041
176	2000	176	0.333	0.333	0.333
176	2000	36	0	1	0
36	2000	36	1	0	0
144	20	144	0.333	0.333	0.333
144	200	176	0.333	0.333	0.333
36	2000	144	1	0	0
36	200	36	1	0	0
176	20	36	0.778	0	0.222
176	200	144	0	1	0
176	20	176	0.333	0	0.667

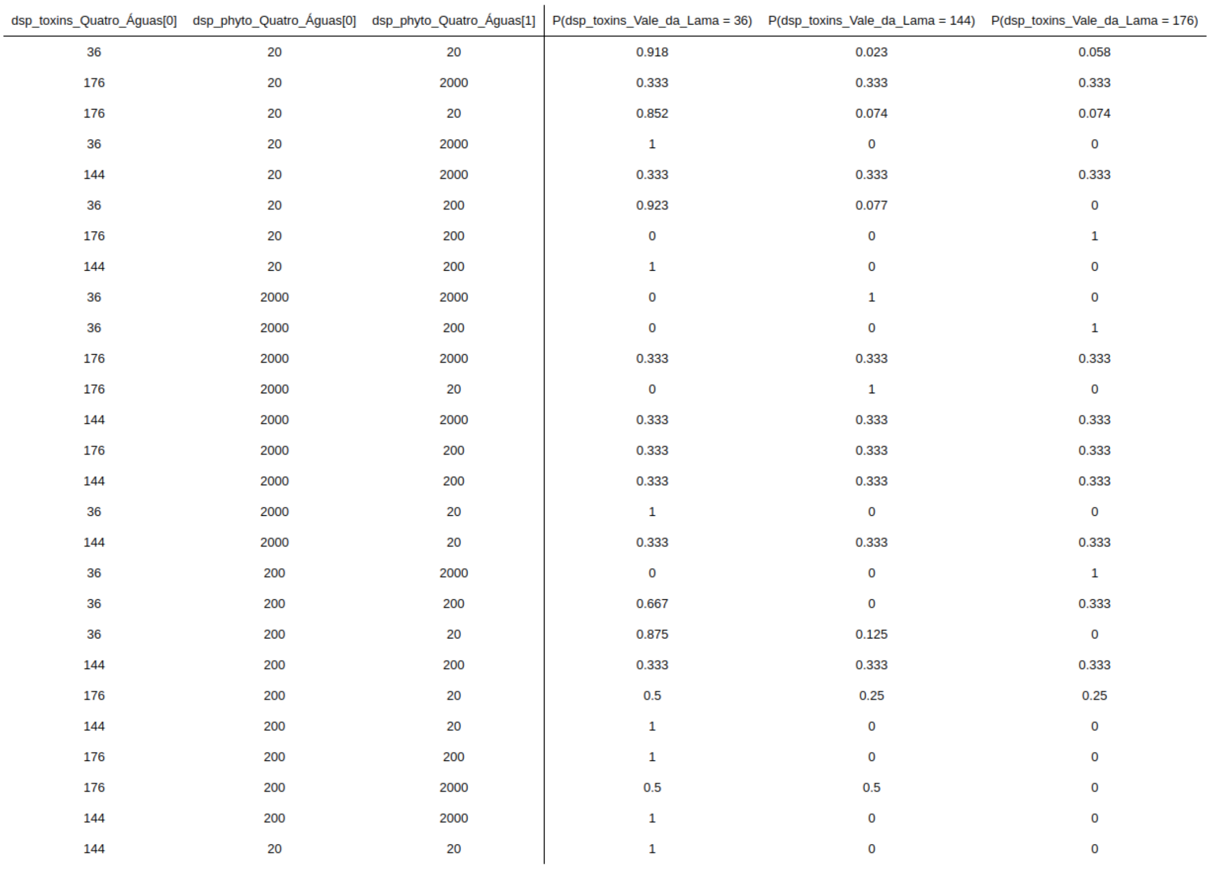
Figure A.3: CPT of DSP Toxins Concentration at OLH2 conditioned in timeslice 1, with pair LAG



dsp_toxins_Fortaleza[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]	P(dsp_phyto_Fortaleza = 20)	P(dsp_phyto_Fortaleza = 200)	P(dsp_phyto_Fortaleza = 2000)
176	200	36	1	0	0
36	200	176	0.333	0.333	0.333
36	200	36	1	0	0
144	200	36	0.333	0.333	0.333
176	200	176	0.333	0.333	0.333
176	2000	176	0.333	0.333	0.333
144	2000	176	0.333	0.333	0.333
36	20	144	0.75	0.25	0
144	20	36	0.9	0	0.1
176	20	176	0.6	0.2	0.2
144	20	176	1	0	0
36	2000	144	1	0	0
36	20	176	0.8	0.1	0.1
176	2000	144	1	0	0
144	2000	144	0.333	0.333	0.333
36	200	144	1	0	0
176	200	144	1	0	0
144	200	144	0.333	0.333	0.333
36	20	36	0.952	0.043	0.005
176	20	36	0.75	0.083	0.167
36	2000	36	0	0	1
176	20	144	1	0	0
144	20	144	0.5	0.5	0
36	2000	176	1	0	0
144	200	176	0.333	0.333	0.333
176	2000	36	1	0	0
144	2000	36	0.333	0.333	0.333

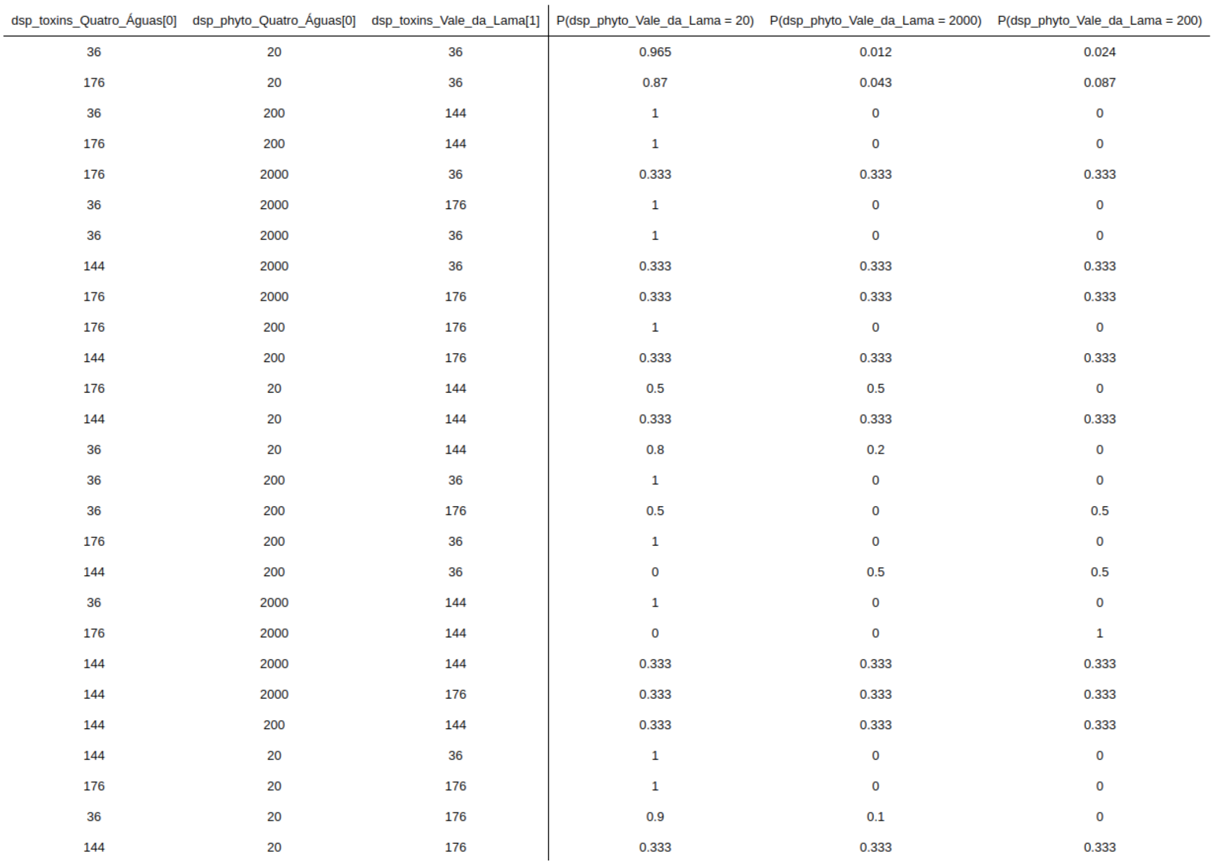
Figure A.4: CPT of DSP Toxins Concentration at OLH2 conditioned in timeslice 1, with pair LAG

A.2 LAG-TAV DBN



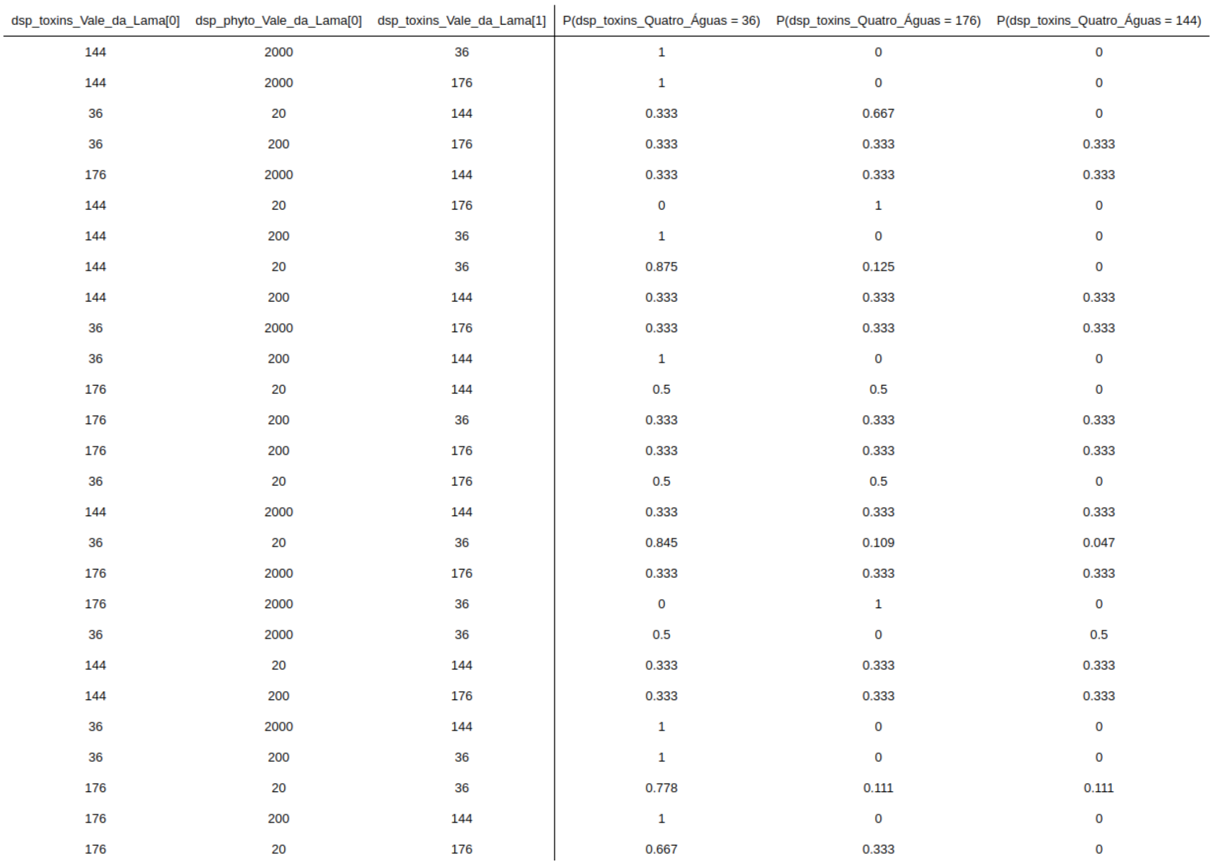
dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]	P(dsp_toxins_Vale_da_Lama = 36)	P(dsp_toxins_Vale_da_Lama = 144)	P(dsp_toxins_Vale_da_Lama = 176)
36	20	20	0.918	0.023	0.058
176	20	2000	0.333	0.333	0.333
176	20	20	0.852	0.074	0.074
36	20	2000	1	0	0
144	20	2000	0.333	0.333	0.333
36	20	200	0.923	0.077	0
176	20	200	0	0	1
144	20	200	1	0	0
36	2000	2000	0	1	0
36	2000	200	0	0	1
176	2000	2000	0.333	0.333	0.333
176	2000	20	0	1	0
144	2000	2000	0.333	0.333	0.333
176	2000	200	0.333	0.333	0.333
144	2000	200	0.333	0.333	0.333
36	2000	20	1	0	0
144	2000	20	0.333	0.333	0.333
36	200	2000	0	0	1
36	200	200	0.667	0	0.333
36	200	20	0.875	0.125	0
144	200	200	0.333	0.333	0.333
176	200	20	0.5	0.25	0.25
144	200	20	1	0	0
176	200	200	1	0	0
176	200	2000	0.5	0.5	0
144	200	2000	1	0	0
144	20	20	1	0	0

Figure A.5: CPT of DSP Toxins Concentration at LAG conditioned in timeslice 1, with pair TAV



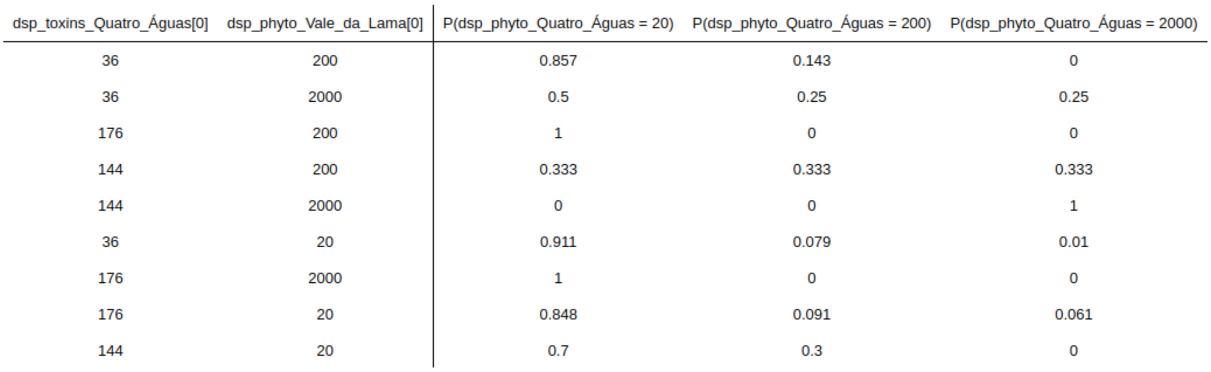
dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_toxins_Vale_da_Lama[1]	P(dsp_phyto_Vale_da_Lama = 20)	P(dsp_phyto_Vale_da_Lama = 2000)	P(dsp_phyto_Vale_da_Lama = 200)
36	20	36	0.965	0.012	0.024
176	20	36	0.87	0.043	0.087
36	200	144	1	0	0
176	200	144	1	0	0
176	2000	36	0.333	0.333	0.333
36	2000	176	1	0	0
36	2000	36	1	0	0
144	2000	36	0.333	0.333	0.333
176	2000	176	0.333	0.333	0.333
176	200	176	1	0	0
144	200	176	0.333	0.333	0.333
176	20	144	0.5	0.5	0
144	20	144	0.333	0.333	0.333
36	20	144	0.8	0.2	0
36	200	36	1	0	0
36	200	176	0.5	0	0.5
176	200	36	1	0	0
144	200	36	0	0.5	0.5
36	2000	144	1	0	0
176	2000	144	0	0	1
144	2000	144	0.333	0.333	0.333
144	2000	176	0.333	0.333	0.333
144	200	144	0.333	0.333	0.333
144	20	36	1	0	0
176	20	176	1	0	0
36	20	176	0.9	0.1	0
144	20	176	0.333	0.333	0.333

Figure A.6: CPT of DSP Phyto Concentration at LAG conditioned in timeslice 1, with pair TAV



dsp_toxins_Vale_da_Lama[0]	dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Vale_da_Lama[1]	P(dsp_toxins_Quatro_Águas = 36)	P(dsp_toxins_Quatro_Águas = 176)	P(dsp_toxins_Quatro_Águas = 144)
144	2000	36	1	0	0
144	2000	176	1	0	0
36	20	144	0.333	0.667	0
36	200	176	0.333	0.333	0.333
176	2000	144	0.333	0.333	0.333
144	20	176	0	1	0
144	200	36	1	0	0
144	20	36	0.875	0.125	0
144	200	144	0.333	0.333	0.333
36	2000	176	0.333	0.333	0.333
36	200	144	1	0	0
176	20	144	0.5	0.5	0
176	200	36	0.333	0.333	0.333
176	200	176	0.333	0.333	0.333
36	20	176	0.5	0.5	0
144	2000	144	0.333	0.333	0.333
36	20	36	0.845	0.109	0.047
176	2000	176	0.333	0.333	0.333
176	2000	36	0	1	0
36	2000	36	0.5	0	0.5
144	20	144	0.333	0.333	0.333
144	200	176	0.333	0.333	0.333
36	2000	144	1	0	0
36	200	36	1	0	0
176	20	36	0.778	0.111	0.111
176	200	144	1	0	0
176	20	176	0.667	0.333	0

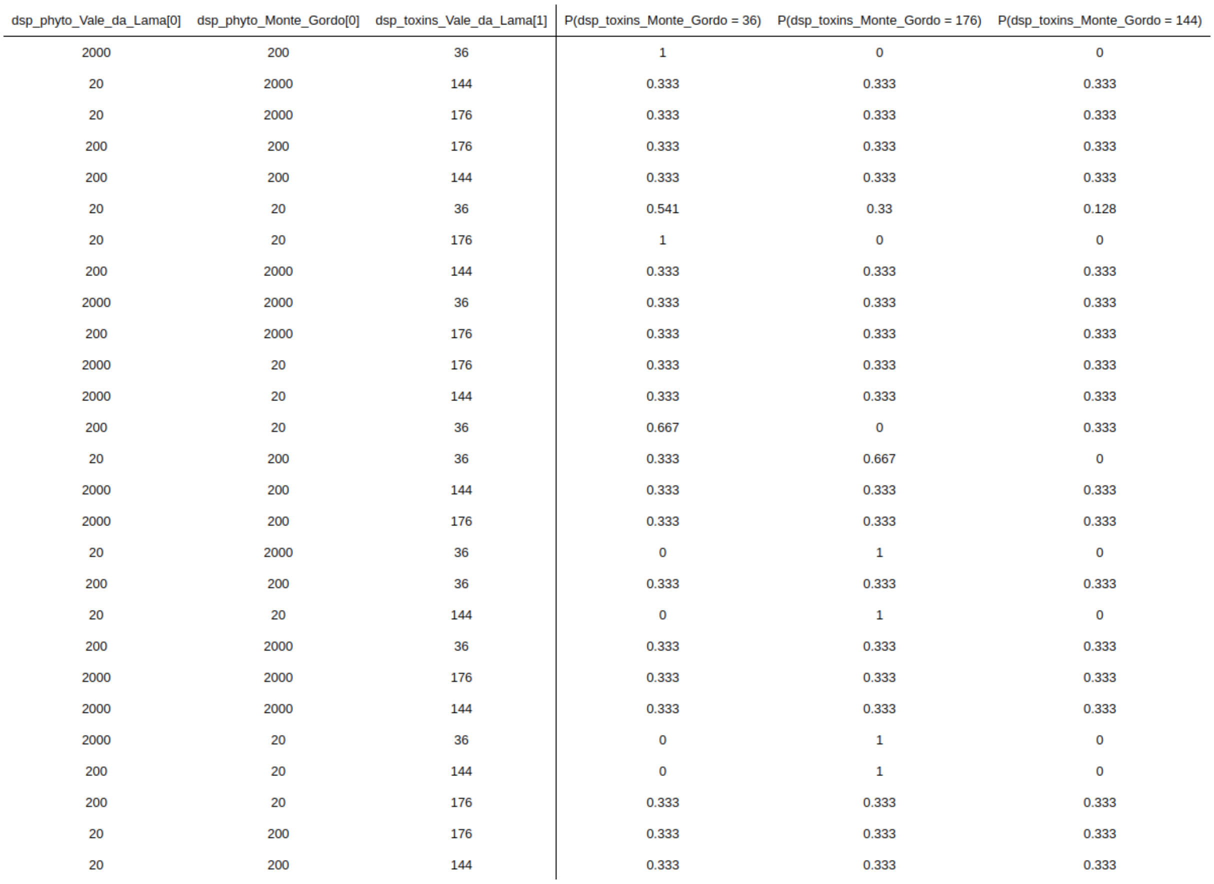
Figure A.7: CPT of DSP Toxins Concentration at TAV conditioned in timeslice 1, with pair LAG



dsp_toxins_Quatro_Águas[0]	dsp_phyto_Vale_da_Lama[0]	P(dsp_phyto_Quatro_Águas = 20)	P(dsp_phyto_Quatro_Águas = 200)	P(dsp_phyto_Quatro_Águas = 2000)
36	200	0.857	0.143	0
36	2000	0.5	0.25	0.25
176	200	1	0	0
144	200	0.333	0.333	0.333
144	2000	0	0	1
36	20	0.911	0.079	0.01
176	2000	1	0	0
176	20	0.848	0.091	0.061
144	20	0.7	0.3	0

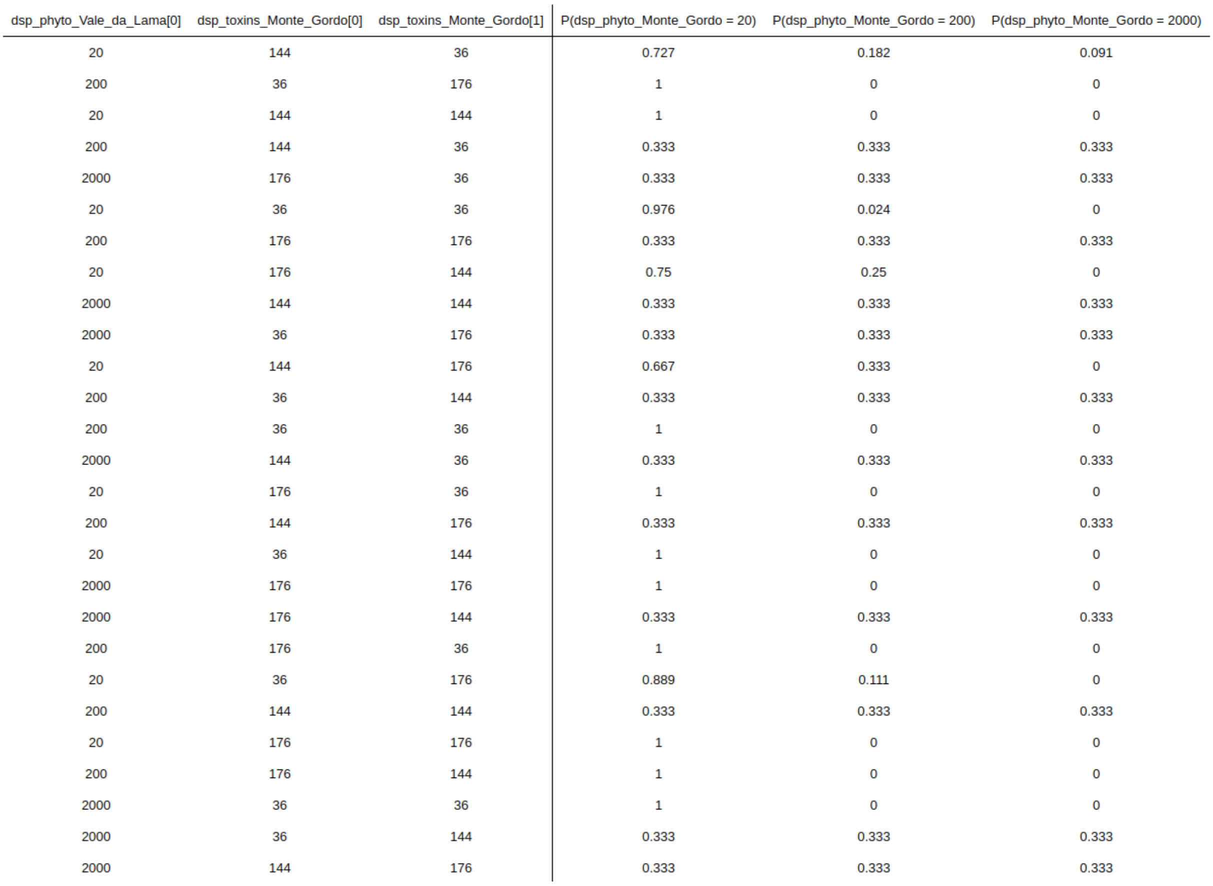
Figure A.8: CPT of DSP Phyto Concentration at TAV conditioned in timeslice 1, with pair LAG

A.3 LAG-L9 DBN



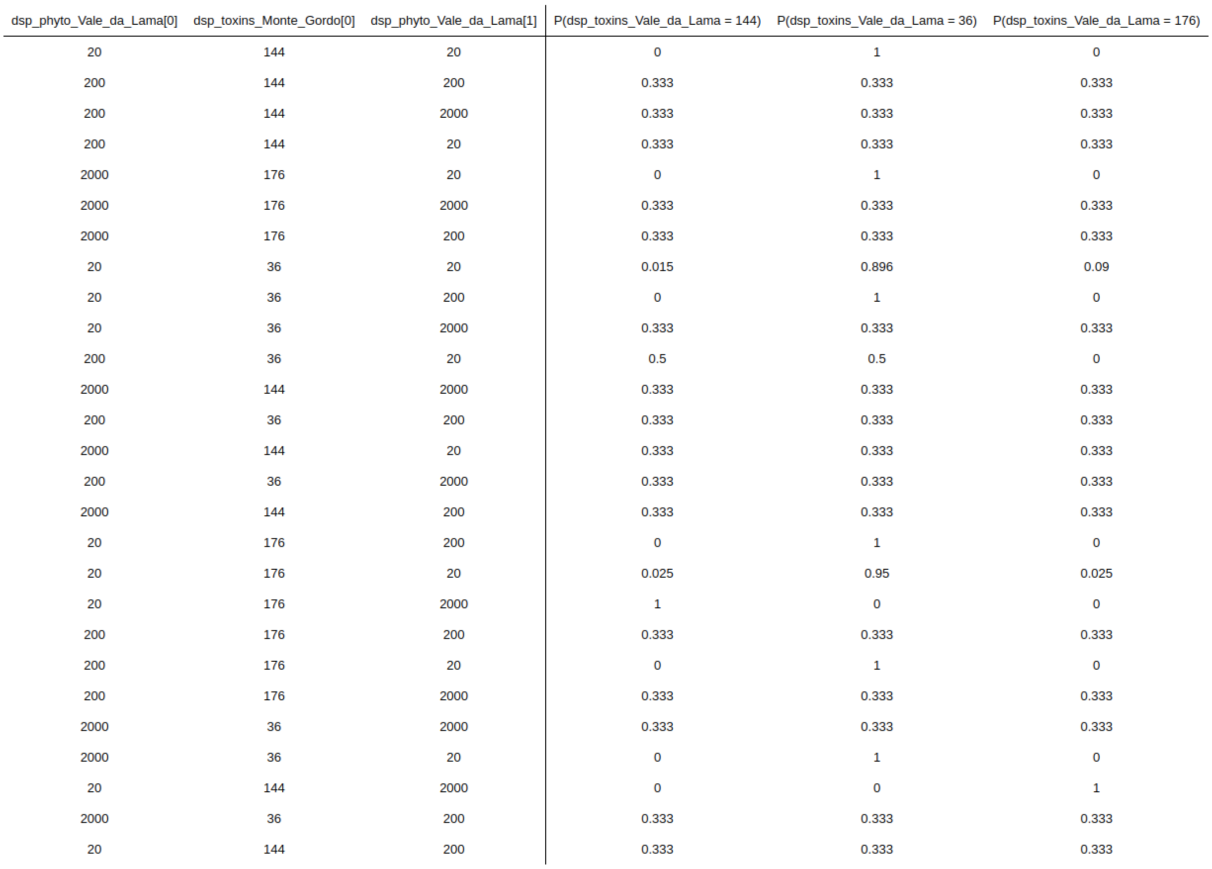
dsp_phyto_Vale_da_Lama[0]	dsp_phyto_Monte_Gordo[0]	dsp_toxins_Vale_da_Lama[1]	P(dsp_toxins_Monte_Gordo = 36)	P(dsp_toxins_Monte_Gordo = 176)	P(dsp_toxins_Monte_Gordo = 144)
2000	200	36	1	0	0
20	2000	144	0.333	0.333	0.333
20	2000	176	0.333	0.333	0.333
200	200	176	0.333	0.333	0.333
200	200	144	0.333	0.333	0.333
20	20	36	0.541	0.33	0.128
20	20	176	1	0	0
200	2000	144	0.333	0.333	0.333
2000	2000	36	0.333	0.333	0.333
200	2000	176	0.333	0.333	0.333
2000	20	176	0.333	0.333	0.333
2000	20	144	0.333	0.333	0.333
200	20	36	0.667	0	0.333
20	200	36	0.333	0.667	0
2000	200	144	0.333	0.333	0.333
2000	200	176	0.333	0.333	0.333
20	2000	36	0	1	0
200	200	36	0.333	0.333	0.333
20	20	144	0	1	0
200	2000	36	0.333	0.333	0.333
2000	2000	176	0.333	0.333	0.333
2000	2000	144	0.333	0.333	0.333
2000	20	36	0	1	0
200	20	144	0	1	0
200	20	176	0.333	0.333	0.333
20	200	176	0.333	0.333	0.333
20	200	144	0.333	0.333	0.333

Figure A.9: CPT of DSP Toxins Concentration at L9 conditioned in timeslice 1, with pair LAG



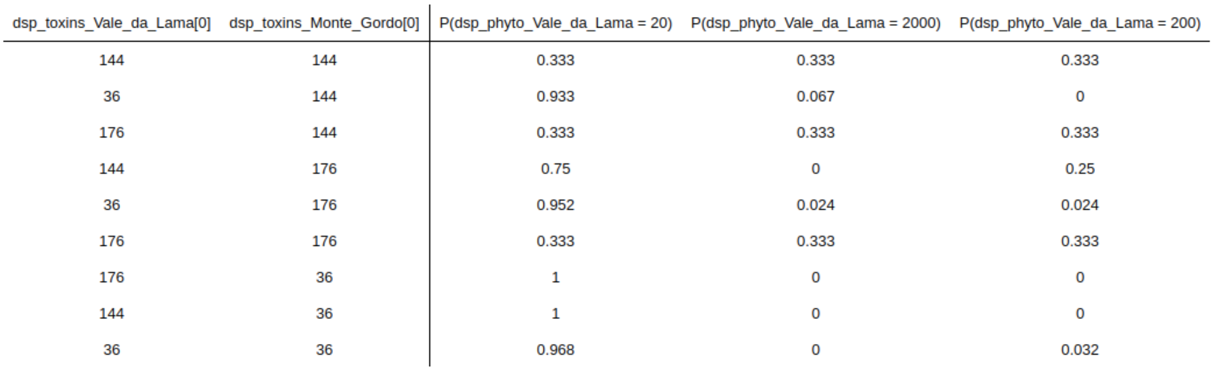
dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Monte_Gordo[0]	dsp_toxins_Monte_Gordo[1]	P(dsp_phyto_Monte_Gordo = 20)	P(dsp_phyto_Monte_Gordo = 200)	P(dsp_phyto_Monte_Gordo = 2000)
20	144	36	0.727	0.182	0.091
200	36	176	1	0	0
20	144	144	1	0	0
200	144	36	0.333	0.333	0.333
2000	176	36	0.333	0.333	0.333
20	36	36	0.976	0.024	0
200	176	176	0.333	0.333	0.333
20	176	144	0.75	0.25	0
2000	144	144	0.333	0.333	0.333
2000	36	176	0.333	0.333	0.333
20	144	176	0.667	0.333	0
200	36	144	0.333	0.333	0.333
200	36	36	1	0	0
2000	144	36	0.333	0.333	0.333
20	176	36	1	0	0
200	144	176	0.333	0.333	0.333
20	36	144	1	0	0
2000	176	176	1	0	0
2000	176	144	0.333	0.333	0.333
200	176	36	1	0	0
20	36	176	0.889	0.111	0
200	144	144	0.333	0.333	0.333
20	176	176	1	0	0
200	176	144	1	0	0
2000	36	36	1	0	0
2000	36	144	0.333	0.333	0.333
2000	144	176	0.333	0.333	0.333

Figure A.10: CPT of DSP Phyto Concentration at L9 conditioned in timeslice 1, with pair LAG



dsp_phyto_Vale_da_Lama[0]	dsp_toxins_Monte_Gordo[0]	dsp_phyto_Vale_da_Lama[1]	P(dsp_toxins_Vale_da_Lama = 144)	P(dsp_toxins_Vale_da_Lama = 36)	P(dsp_toxins_Vale_da_Lama = 176)
20	144	20	0	1	0
200	144	200	0.333	0.333	0.333
200	144	2000	0.333	0.333	0.333
200	144	20	0.333	0.333	0.333
2000	176	20	0	1	0
2000	176	2000	0.333	0.333	0.333
2000	176	200	0.333	0.333	0.333
20	36	20	0.015	0.896	0.09
20	36	200	0	1	0
20	36	2000	0.333	0.333	0.333
200	36	20	0.5	0.5	0
2000	144	2000	0.333	0.333	0.333
200	36	200	0.333	0.333	0.333
2000	144	20	0.333	0.333	0.333
200	36	2000	0.333	0.333	0.333
2000	144	200	0.333	0.333	0.333
20	176	200	0	1	0
20	176	20	0.025	0.95	0.025
20	176	2000	1	0	0
200	176	200	0.333	0.333	0.333
200	176	20	0	1	0
200	176	2000	0.333	0.333	0.333
2000	36	2000	0.333	0.333	0.333
2000	36	20	0	1	0
20	144	2000	0	0	1
2000	36	200	0.333	0.333	0.333
20	144	200	0.333	0.333	0.333

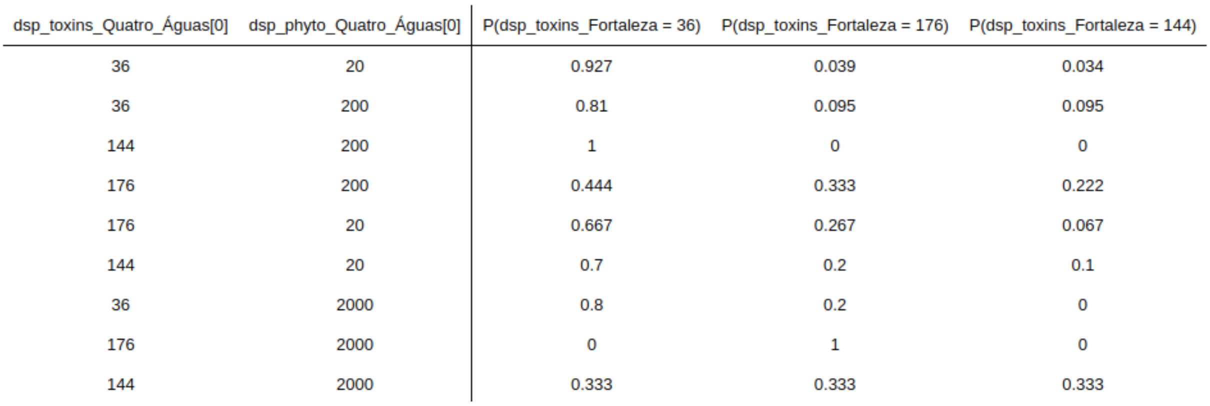
Figure A.11: CPT of DSP Toxins Concentration at LAG conditioned in timeslice 1, with pair L9



dsp_toxins_Vale_da_Lama[0]	dsp_toxins_Monte_Gordo[0]	P(dsp_phyto_Vale_da_Lama = 20)	P(dsp_phyto_Vale_da_Lama = 2000)	P(dsp_phyto_Vale_da_Lama = 200)
144	144	0.333	0.333	0.333
36	144	0.933	0.067	0
176	144	0.333	0.333	0.333
144	176	0.75	0	0.25
36	176	0.952	0.024	0.024
176	176	0.333	0.333	0.333
176	36	1	0	0
144	36	1	0	0
36	36	0.968	0	0.032

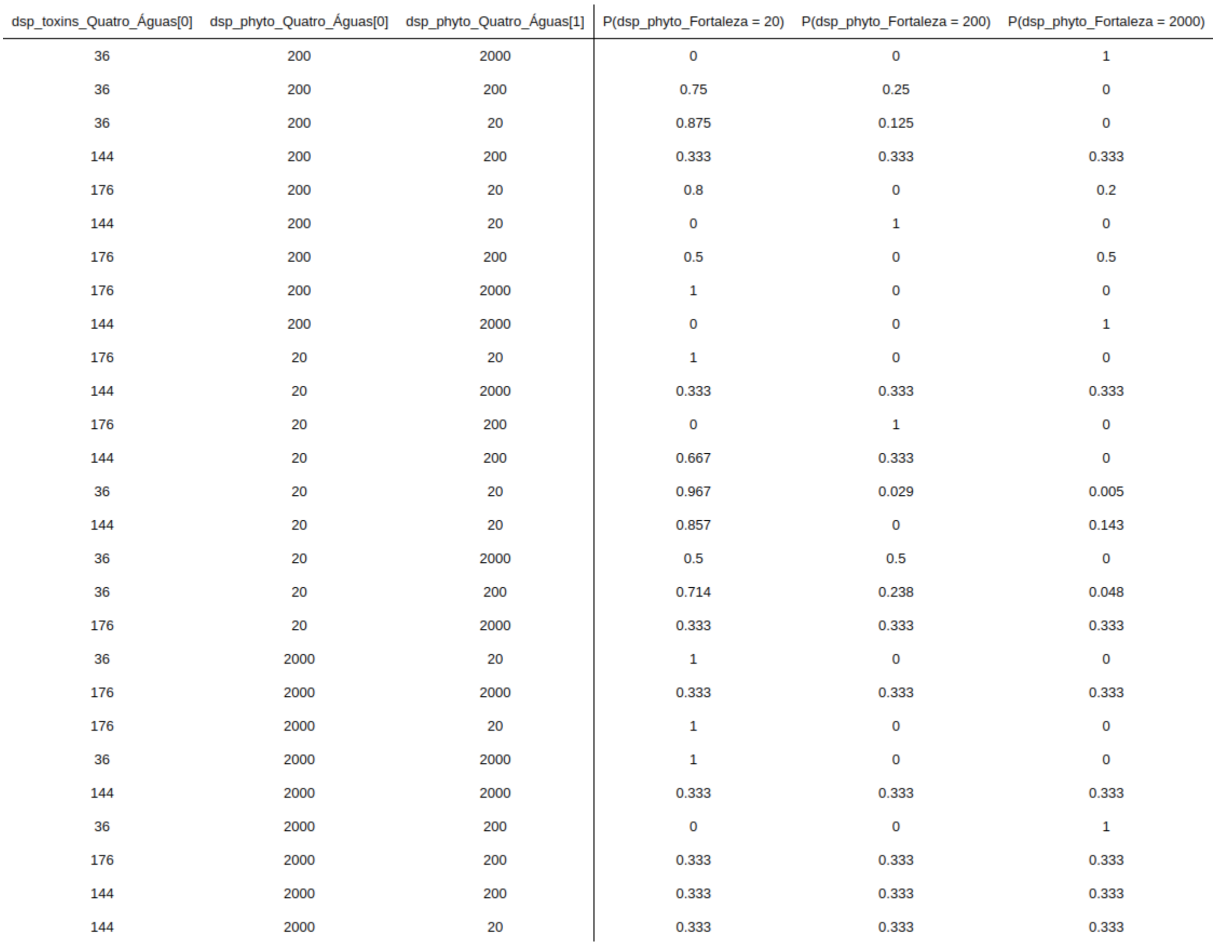
Figure A.12: CPT of DSP Phyto Concentration at LAG conditioned in timeslice 1, with pair L9

A.4 OLH2-TAV DBN



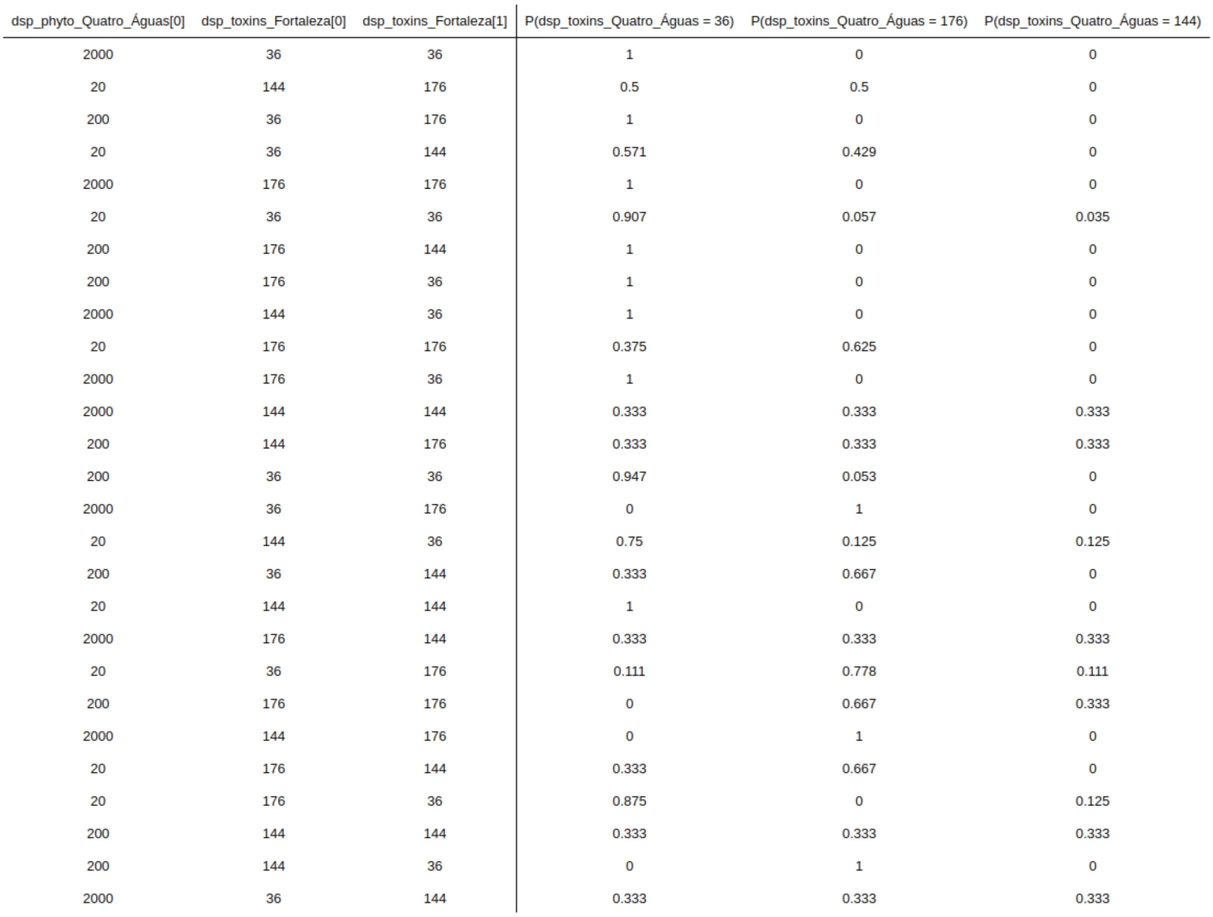
dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	P(dsp_toxins_Fortaleza = 36)	P(dsp_toxins_Fortaleza = 176)	P(dsp_toxins_Fortaleza = 144)
36	20	0.927	0.039	0.034
36	200	0.81	0.095	0.095
144	200	1	0	0
176	200	0.444	0.333	0.222
176	20	0.667	0.267	0.067
144	20	0.7	0.2	0.1
36	2000	0.8	0.2	0
176	2000	0	1	0
144	2000	0.333	0.333	0.333

Figure A.13: CPT of DSP Toxins Concentration at OLH2 conditioned in timeslice 1, with pair TAV



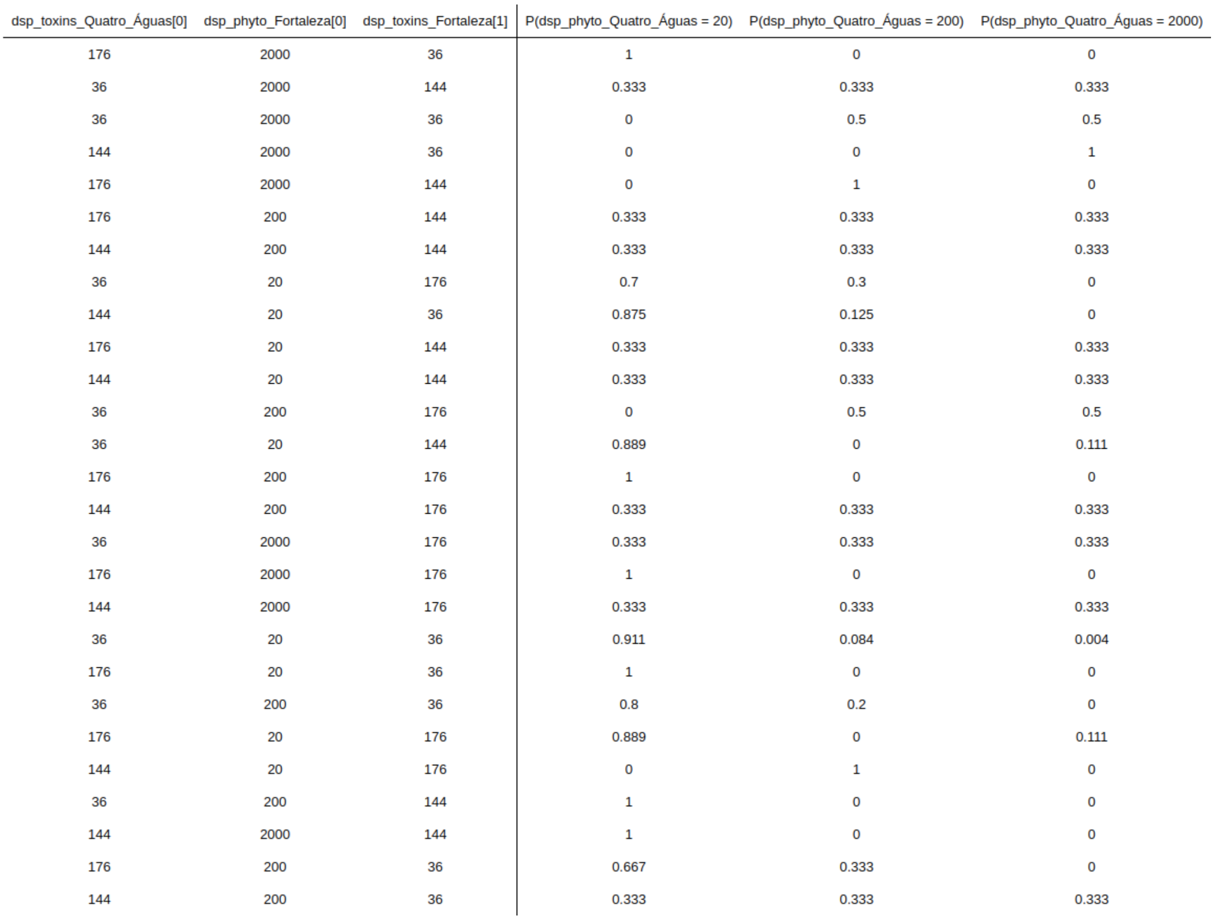
dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]	P(dsp_phyto_Fortaleza = 20)	P(dsp_phyto_Fortaleza = 200)	P(dsp_phyto_Fortaleza = 2000)
36	200	2000	0	0	1
36	200	200	0.75	0.25	0
36	200	20	0.875	0.125	0
144	200	200	0.333	0.333	0.333
176	200	20	0.8	0	0.2
144	200	20	0	1	0
176	200	200	0.5	0	0.5
176	200	2000	1	0	0
144	200	2000	0	0	1
176	20	20	1	0	0
144	20	2000	0.333	0.333	0.333
176	20	200	0	1	0
144	20	200	0.667	0.333	0
36	20	20	0.967	0.029	0.005
144	20	20	0.857	0	0.143
36	20	2000	0.5	0.5	0
36	20	200	0.714	0.238	0.048
176	20	2000	0.333	0.333	0.333
36	2000	20	1	0	0
176	2000	2000	0.333	0.333	0.333
176	2000	20	1	0	0
36	2000	2000	1	0	0
144	2000	2000	0.333	0.333	0.333
36	2000	200	0	0	1
176	2000	200	0.333	0.333	0.333
144	2000	200	0.333	0.333	0.333
144	2000	20	0.333	0.333	0.333

Figure A.14: CPT of DSP Phyto Concentration at OLH2 conditioned in timeslice 1, with pair TAV



dsp_phyto_Quatro_Águas[0]	dsp_toxins_Fortaleza[0]	dsp_toxins_Fortaleza[1]	P(dsp_toxins_Quatro_Águas = 36)	P(dsp_toxins_Quatro_Águas = 176)	P(dsp_toxins_Quatro_Águas = 144)
2000	36	36	1	0	0
20	144	176	0.5	0.5	0
200	36	176	1	0	0
20	36	144	0.571	0.429	0
2000	176	176	1	0	0
20	36	36	0.907	0.057	0.035
200	176	144	1	0	0
200	176	36	1	0	0
2000	144	36	1	0	0
20	176	176	0.375	0.625	0
2000	176	36	1	0	0
2000	144	144	0.333	0.333	0.333
200	144	176	0.333	0.333	0.333
200	36	36	0.947	0.053	0
2000	36	176	0	1	0
20	144	36	0.75	0.125	0.125
200	36	144	0.333	0.667	0
20	144	144	1	0	0
2000	176	144	0.333	0.333	0.333
20	36	176	0.111	0.778	0.111
200	176	176	0	0.667	0.333
2000	144	176	0	1	0
20	176	144	0.333	0.667	0
20	176	36	0.875	0	0.125
200	144	144	0.333	0.333	0.333
200	144	36	0	1	0
2000	36	144	0.333	0.333	0.333

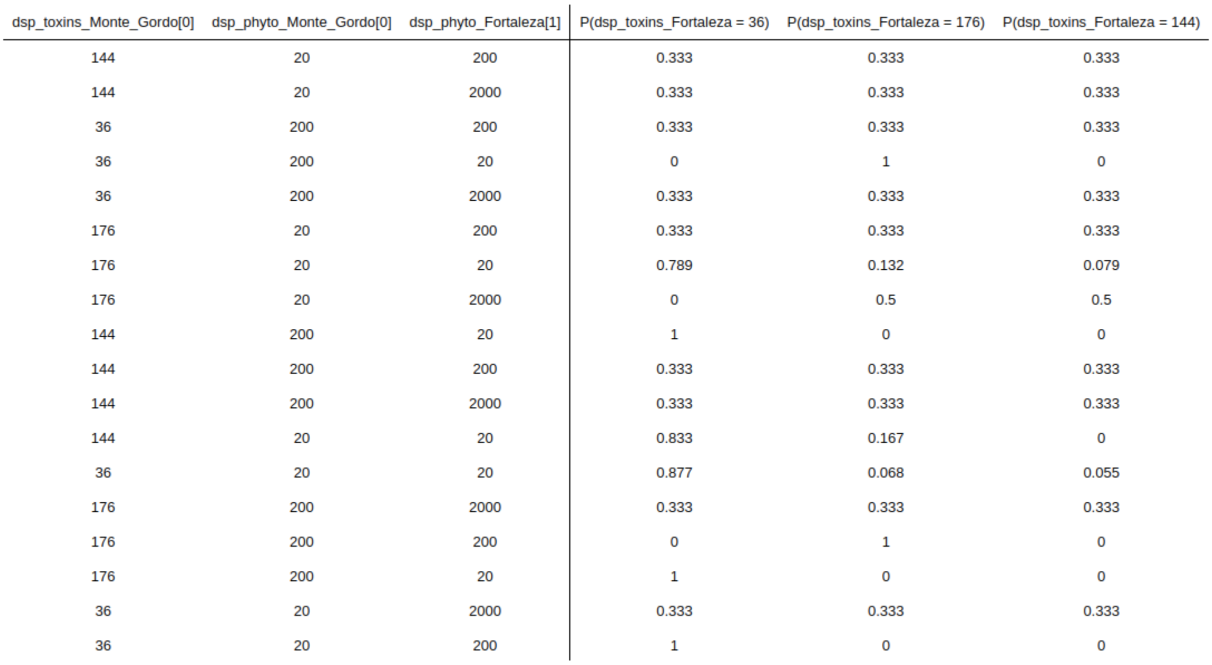
Figure A.15: CPT of DSP Toxins Concentration at TAV conditioned in timeslice 1, with pair OLH2



dsp_toxins_Quatro_Águas[0]	dsp_phyto_Fortaleza[0]	dsp_toxins_Fortaleza[1]	P(dsp_phyto_Quatro_Águas = 20)	P(dsp_phyto_Quatro_Águas = 200)	P(dsp_phyto_Quatro_Águas = 2000)
176	2000	36	1	0	0
36	2000	144	0.333	0.333	0.333
36	2000	36	0	0.5	0.5
144	2000	36	0	0	1
176	2000	144	0	1	0
176	200	144	0.333	0.333	0.333
144	200	144	0.333	0.333	0.333
36	20	176	0.7	0.3	0
144	20	36	0.875	0.125	0
176	20	144	0.333	0.333	0.333
144	20	144	0.333	0.333	0.333
36	200	176	0	0.5	0.5
36	20	144	0.889	0	0.111
176	200	176	1	0	0
144	200	176	0.333	0.333	0.333
36	2000	176	0.333	0.333	0.333
176	2000	176	1	0	0
144	2000	176	0.333	0.333	0.333
36	20	36	0.911	0.084	0.004
176	20	36	1	0	0
36	200	36	0.8	0.2	0
176	20	176	0.889	0	0.111
144	20	176	0	1	0
36	200	144	1	0	0
144	2000	144	1	0	0
176	200	36	0.667	0.333	0
144	200	36	0.333	0.333	0.333

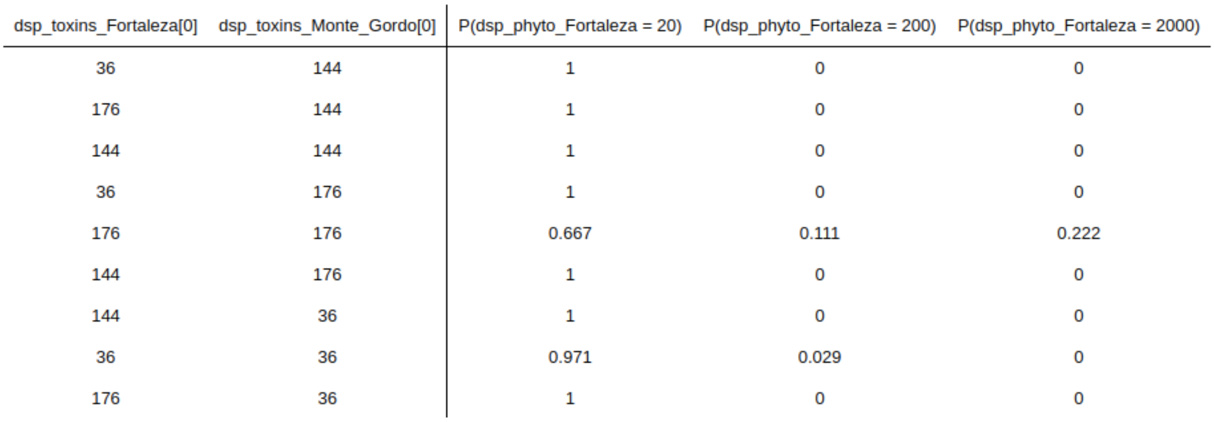
Figure A.16: CPT of DSP Phyto Concentration at TAV conditioned in timeslice 1, with pair OLH2

A.5 OLH2-L9 DBN



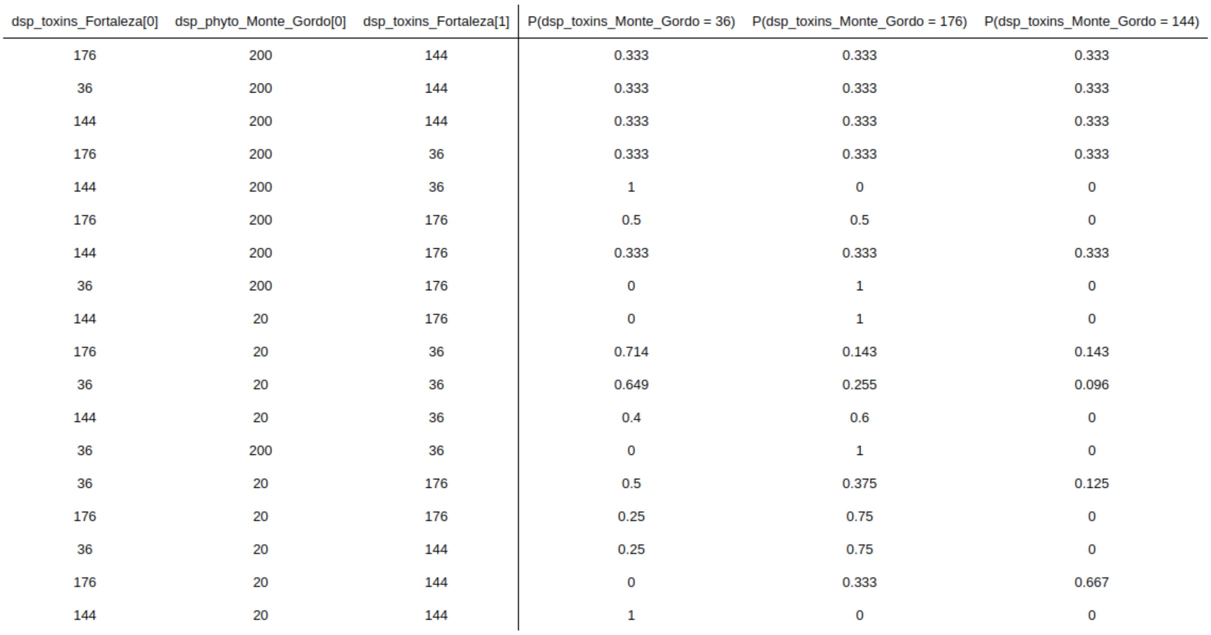
dsp_toxins_Monte_Gordo[0]	dsp_phyto_Monte_Gordo[0]	dsp_phyto_Fortaleza[1]	P(dsp_toxins_Fortaleza = 36)	P(dsp_toxins_Fortaleza = 176)	P(dsp_toxins_Fortaleza = 144)
144	20	200	0.333	0.333	0.333
144	20	2000	0.333	0.333	0.333
36	200	200	0.333	0.333	0.333
36	200	20	0	1	0
36	200	2000	0.333	0.333	0.333
176	20	200	0.333	0.333	0.333
176	20	20	0.789	0.132	0.079
176	20	2000	0	0.5	0.5
144	200	20	1	0	0
144	200	200	0.333	0.333	0.333
144	200	2000	0.333	0.333	0.333
144	20	20	0.833	0.167	0
36	20	20	0.877	0.068	0.055
176	200	2000	0.333	0.333	0.333
176	200	200	0	1	0
176	200	20	1	0	0
36	20	2000	0.333	0.333	0.333
36	20	200	1	0	0

Figure A.17: CPT of DSP Toxins Concentration at OLH2 conditioned in timeslice 1, with pair L9



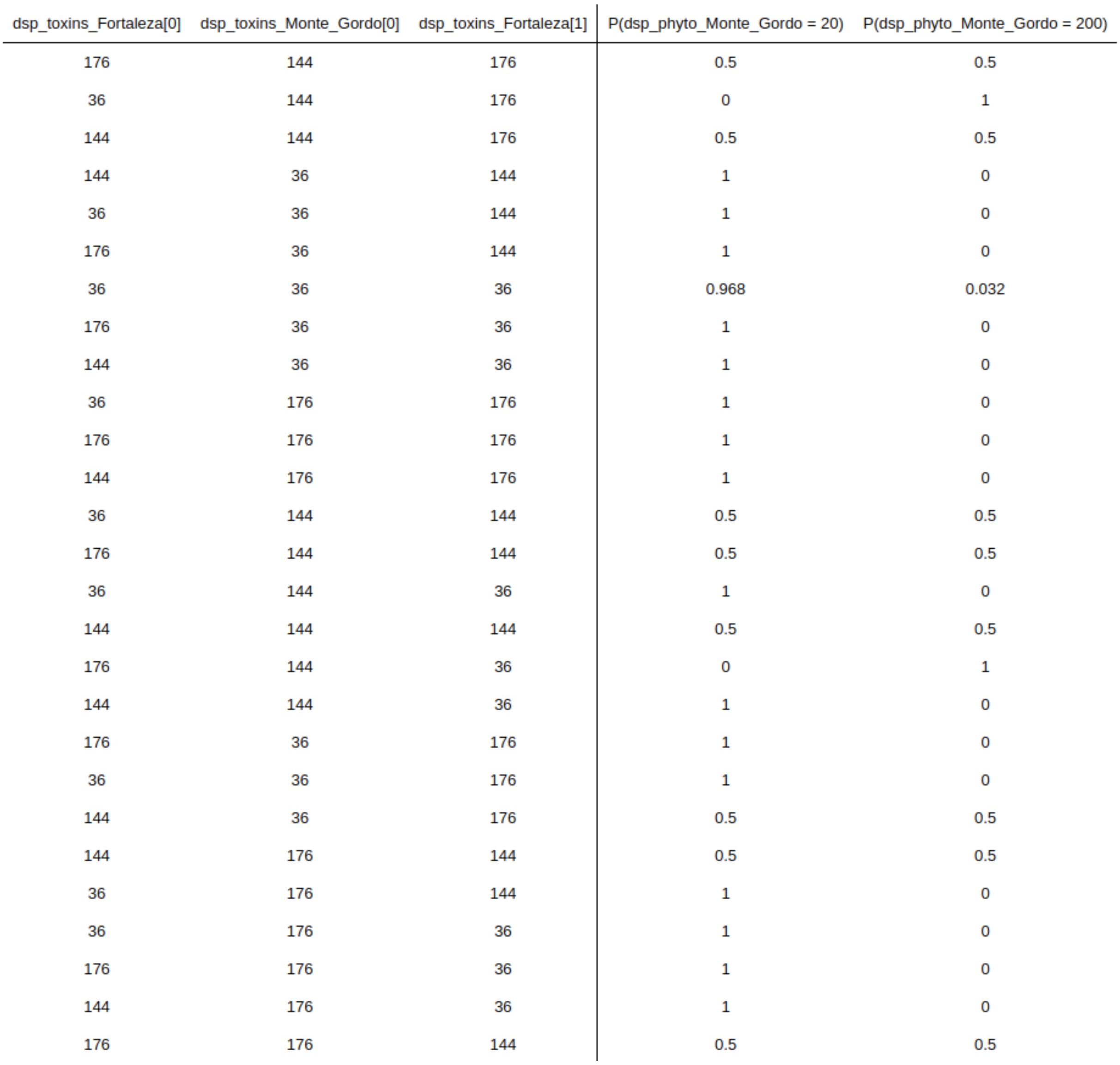
dsp_toxins_Fortaleza[0]	dsp_toxins_Monte_Gordo[0]	P(dsp_phyto_Fortaleza = 20)	P(dsp_phyto_Fortaleza = 200)	P(dsp_phyto_Fortaleza = 2000)
36	144	1	0	0
176	144	1	0	0
144	144	1	0	0
36	176	1	0	0
176	176	0.667	0.111	0.222
144	176	1	0	0
144	36	1	0	0
36	36	0.971	0.029	0
176	36	1	0	0

Figure A.18: CPT of DSP Phyto Concentration at OLH2 conditioned in timeslice 1, with pair L9



dsp_toxins_Fortaleza[0]	dsp_phyto_Monte_Gordo[0]	dsp_toxins_Fortaleza[1]	P(dsp_toxins_Monte_Gordo = 36)	P(dsp_toxins_Monte_Gordo = 176)	P(dsp_toxins_Monte_Gordo = 144)
176	200	144	0.333	0.333	0.333
36	200	144	0.333	0.333	0.333
144	200	144	0.333	0.333	0.333
176	200	36	0.333	0.333	0.333
144	200	36	1	0	0
176	200	176	0.5	0.5	0
144	200	176	0.333	0.333	0.333
36	200	176	0	1	0
144	20	176	0	1	0
176	20	36	0.714	0.143	0.143
36	20	36	0.649	0.255	0.096
144	20	36	0.4	0.6	0
36	200	36	0	1	0
36	20	176	0.5	0.375	0.125
176	20	176	0.25	0.75	0
36	20	144	0.25	0.75	0
176	20	144	0	0.333	0.667
144	20	144	1	0	0

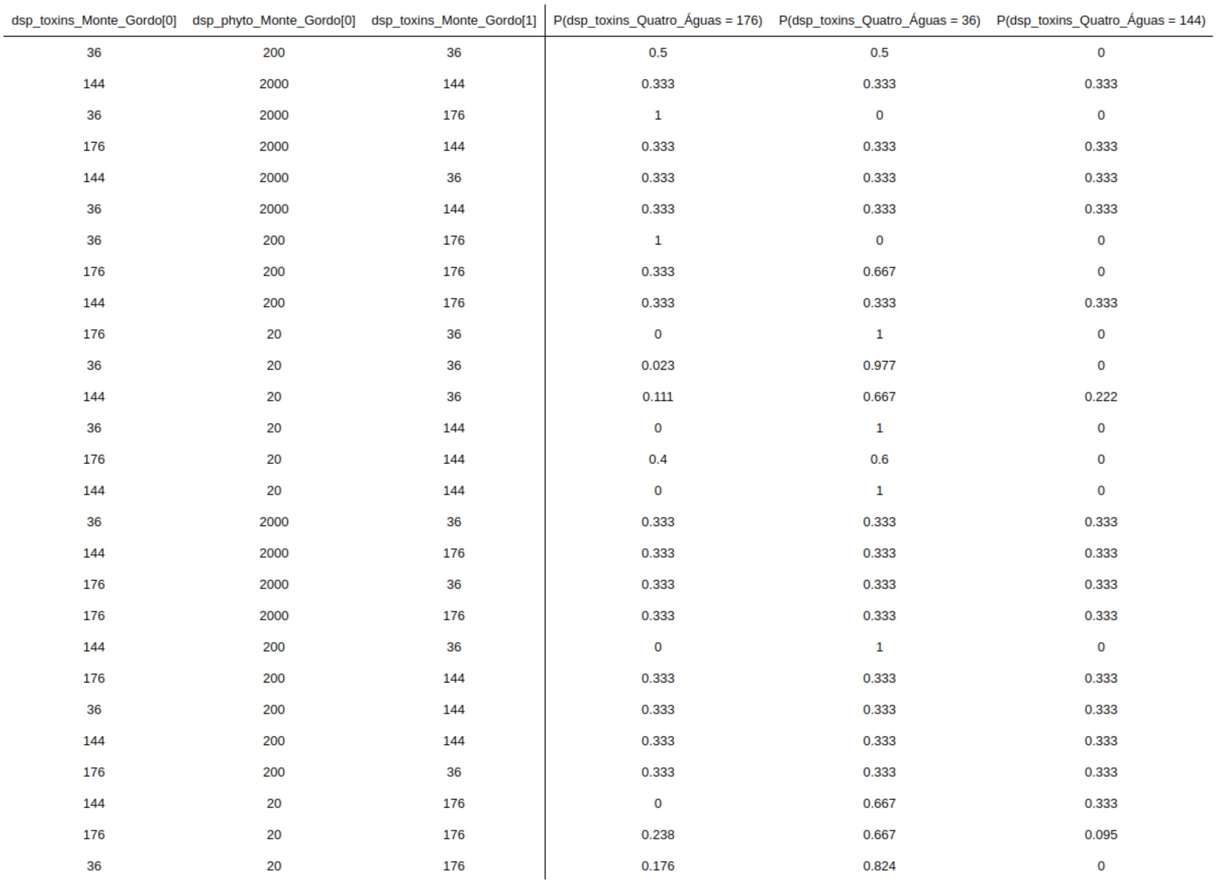
Figure A.19: CPT of DSP Toxins Concentration at L9 conditioned in timeslice 1, with pair OLH2



dsp_toxins_Fortaleza[0]	dsp_toxins_Monte_Gordo[0]	dsp_toxins_Fortaleza[1]	P(dsp_phyto_Monte_Gordo = 20)	P(dsp_phyto_Monte_Gordo = 200)
176	144	176	0.5	0.5
36	144	176	0	1
144	144	176	0.5	0.5
144	36	144	1	0
36	36	144	1	0
176	36	144	1	0
36	36	36	0.968	0.032
176	36	36	1	0
144	36	36	1	0
36	176	176	1	0
176	176	176	1	0
144	176	176	1	0
36	144	144	0.5	0.5
176	144	144	0.5	0.5
36	144	36	1	0
144	144	144	0.5	0.5
176	144	36	0	1
144	144	36	1	0
176	36	176	1	0
36	36	176	1	0
144	36	176	0.5	0.5
144	176	144	0.5	0.5
36	176	144	1	0
36	176	36	1	0
176	176	36	1	0
144	176	36	1	0
176	176	144	0.5	0.5

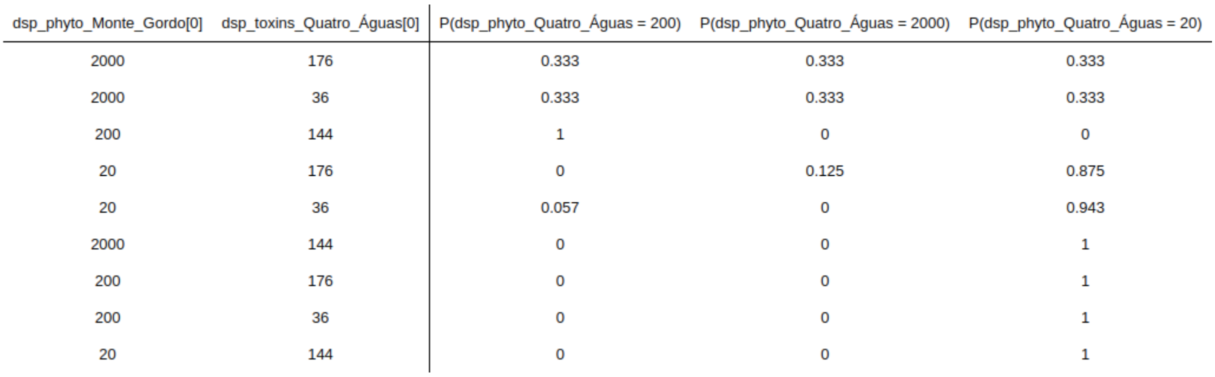
Figure A.20: CPT of DSP Phyto Concentration at L9 conditioned in timeslice 1, with pair OLH2

A.6 TAV-L9 DBN



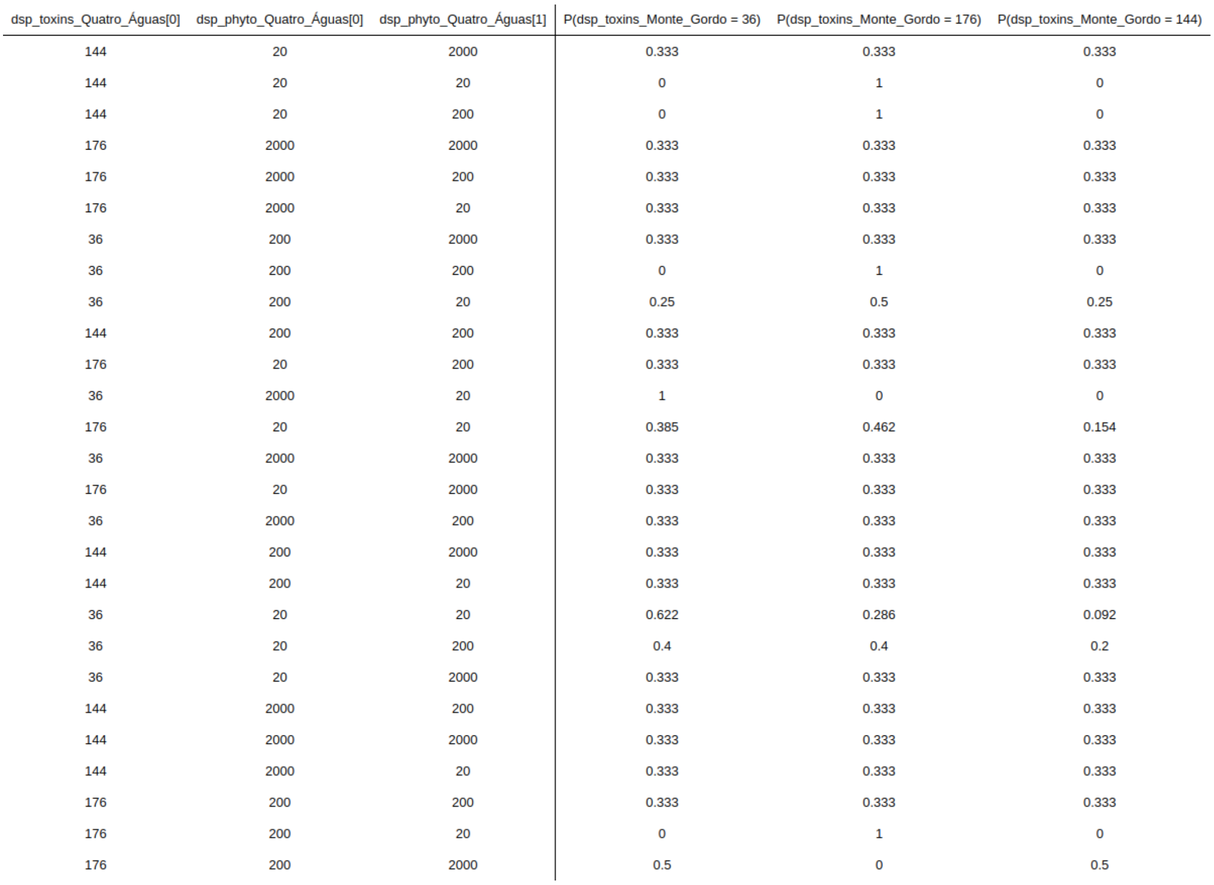
dsp_toxins_Monte_Gordo[0]	dsp_phyto_Monte_Gordo[0]	dsp_toxins_Monte_Gordo[1]	P(dsp_toxins_Quatro_Águas = 176)	P(dsp_toxins_Quatro_Águas = 36)	P(dsp_toxins_Quatro_Águas = 144)
36	200	36	0.5	0.5	0
144	2000	144	0.333	0.333	0.333
36	2000	176	1	0	0
176	2000	144	0.333	0.333	0.333
144	2000	36	0.333	0.333	0.333
36	2000	144	0.333	0.333	0.333
36	200	176	1	0	0
176	200	176	0.333	0.667	0
144	200	176	0.333	0.333	0.333
176	20	36	0	1	0
36	20	36	0.023	0.977	0
144	20	36	0.111	0.667	0.222
36	20	144	0	1	0
176	20	144	0.4	0.6	0
144	20	144	0	1	0
36	2000	36	0.333	0.333	0.333
144	2000	176	0.333	0.333	0.333
176	2000	36	0.333	0.333	0.333
176	2000	176	0.333	0.333	0.333
144	200	36	0	1	0
176	200	144	0.333	0.333	0.333
36	200	144	0.333	0.333	0.333
144	200	144	0.333	0.333	0.333
176	200	36	0.333	0.333	0.333
144	20	176	0	0.667	0.333
176	20	176	0.238	0.667	0.095
36	20	176	0.176	0.824	0

Figure A.21: CPT of DSP Toxins Concentration at TAV conditioned in timeslice 1, with pair L9



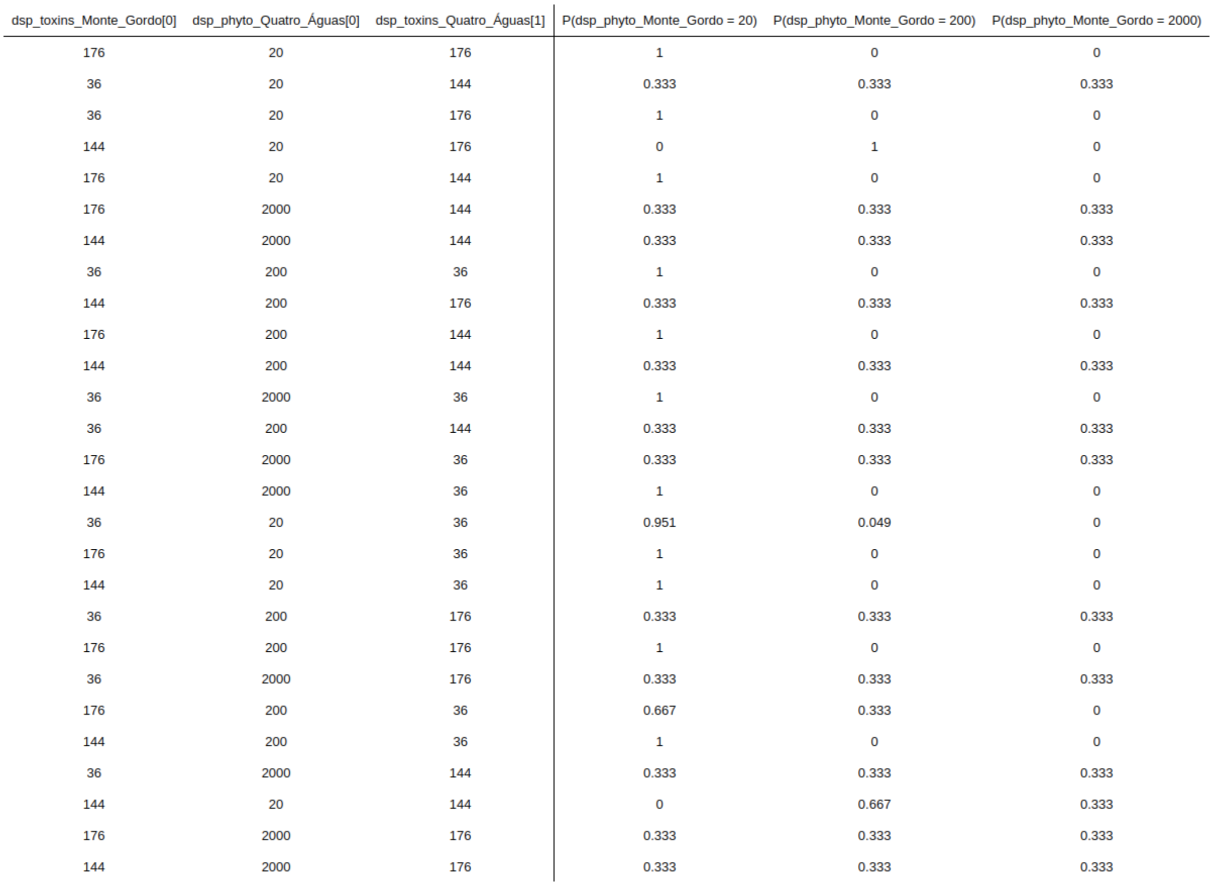
dsp_phyto_Monte_Gordo[0]	dsp_toxins_Quatro_Águas[0]	P(dsp_phyto_Quatro_Águas = 200)	P(dsp_phyto_Quatro_Águas = 2000)	P(dsp_phyto_Quatro_Águas = 20)
2000	176	0.333	0.333	0.333
2000	36	0.333	0.333	0.333
200	144	1	0	0
20	176	0	0.125	0.875
20	36	0.057	0	0.943
2000	144	0	0	1
200	176	0	0	1
200	36	0	0	1
20	144	0	0	1

Figure A.22: CPT of DSP Phyto Concentration at TAV conditioned in timeslice 1, with pair L9



dsp_toxins_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[0]	dsp_phyto_Quatro_Águas[1]	P(dsp_toxins_Monte_Gordo = 36)	P(dsp_toxins_Monte_Gordo = 176)	P(dsp_toxins_Monte_Gordo = 144)
144	20	2000	0.333	0.333	0.333
144	20	20	0	1	0
144	20	200	0	1	0
176	2000	2000	0.333	0.333	0.333
176	2000	200	0.333	0.333	0.333
176	2000	20	0.333	0.333	0.333
36	200	2000	0.333	0.333	0.333
36	200	200	0	1	0
36	200	20	0.25	0.5	0.25
144	200	200	0.333	0.333	0.333
176	20	200	0.333	0.333	0.333
36	2000	20	1	0	0
176	20	20	0.385	0.462	0.154
36	2000	2000	0.333	0.333	0.333
176	20	2000	0.333	0.333	0.333
36	2000	200	0.333	0.333	0.333
144	200	2000	0.333	0.333	0.333
144	200	20	0.333	0.333	0.333
36	20	20	0.622	0.286	0.092
36	20	200	0.4	0.4	0.2
36	20	2000	0.333	0.333	0.333
144	2000	200	0.333	0.333	0.333
144	2000	2000	0.333	0.333	0.333
144	2000	20	0.333	0.333	0.333
176	200	200	0.333	0.333	0.333
176	200	20	0	1	0
176	200	2000	0.5	0	0.5

Figure A.23: CPT of DSP Toxins Concentration at L9 conditioned in timeslice 1, with pair TAV



dsp_toxins_Monte_Gordo[0]	dsp_phyto_Quatro_Águas[0]	dsp_toxins_Quatro_Águas[1]	P(dsp_phyto_Monte_Gordo = 20)	P(dsp_phyto_Monte_Gordo = 200)	P(dsp_phyto_Monte_Gordo = 2000)
176	20	176	1	0	0
36	20	144	0.333	0.333	0.333
36	20	176	1	0	0
144	20	176	0	1	0
176	20	144	1	0	0
176	2000	144	0.333	0.333	0.333
144	2000	144	0.333	0.333	0.333
36	200	36	1	0	0
144	200	176	0.333	0.333	0.333
176	200	144	1	0	0
144	200	144	0.333	0.333	0.333
36	2000	36	1	0	0
36	200	144	0.333	0.333	0.333
176	2000	36	0.333	0.333	0.333
144	2000	36	1	0	0
36	20	36	0.951	0.049	0
176	20	36	1	0	0
144	20	36	1	0	0
36	200	176	0.333	0.333	0.333
176	200	176	1	0	0
36	2000	176	0.333	0.333	0.333
176	200	36	0.667	0.333	0
144	200	36	1	0	0
36	2000	144	0.333	0.333	0.333
144	20	144	0	0.667	0.333
176	2000	176	0.333	0.333	0.333
144	2000	176	0.333	0.333	0.333

Figure A.24: CPT of DSP Phyto Concentration at L9 conditioned in timeslice 1, with pair TAV