Causal graph discovery for explainable insights on marine biotoxin shellfish contamination

Diogo Ribeiro, Susana Vinga and Alexandra M. Carvalho

***Abstract*—Harmful algal blooms are natural phenomena that cause shellfish contamination due to the rapid accumulation of marine biotoxins. The Portuguese Institute of the Ocean and the Atmosphere (IPMA) regularly monitors toxic phytoplankton and temporarily closes shellfish production areas whenever biotoxin concentration exceeds safety limits to prevent public health risks. Causality techniques applied to multivariate time series data can identify the variables that most influence marine biotoxin contamination. In this way, several state-of-the-art methods were validated in benchmark datasets. The Peter-Clark Momentary Conditional Independence plus (PCMCI+) parameterization using Gaussian Processes and Distance Correlation conditional independence test (GPDC) with $\alpha_{\mathrm{PC}} = 0.3$ stood out. Based on this algorithm, it was explored the biotoxin concentration in mussels *Mytilus galloprovincialis*, donax clams *Donax trunculus*, cockles *Cerastoderma edule* and environmental data from IPMA and Copernicus Marine Environment Monitoring Service (CMEMS). We conclude that maximum temperature, wind intensity and rainfall are predictors of mussel contamination with DSP toxins for shorter-term dependencies and *chl-a* for longer-term dependencies. Concerning donax clams contamination, only the maximum air temperature with a 1-week temporal lag was inferred. Cockle contamination showed a great connection with biological variables, namely DSP toxins-producing phytoplankton and DSP toxins. Additionally, a maximum air temperature manifests with a temporal lag of 1-week and chl-a with a temporal lag of 3-weeks. This study proposes a novel approach to infer the relationships between environmental variables to enhance decision-making and public health safety regarding shellfish consumption in Portugal.**

***Index Terms*—Multivariate Time Series, Shellfish Contamination, Marine Biotoxins, Public Health, Causal Models, Causal Discovery**

I. Introduction

A. Topic Overview

Aquaculture plays a crucial role in satisfying the demand for seafood production, namely bivalve mollusc production. This critical activity worldwide provides healthy food and essential economic support for remote families and small businesses in rural areas with limited job opportunities. In Portugal, the production and harvest of molluscan shellfish have increased from the north to the south coast, with a significant impact on the national economy [1].

From the consumer's point of view, since bivalve molluscs are filter-feeding organisms, they are a healthy product that feeds on natural phytoplankton in the water column, without the typical fertilisers used in land-based agriculture .However, some microalgae species produce high concentrations of biotoxins that can accumulate and contaminate shellfish, making them unsafe for human consumption. This event is known as Harmful Algal Blooms (HABs) and is described by the rapid and uncontrolled growth of toxic phytoplankton when environmental conditions (such as atmospheric, oceanographic and biological) are favourable.

TABLE I: Limits regulated by the European Commission. Adapted from [2].

Marine biotoxins	Regulated Limits
PSP (Paralytic Shellfish Poison)	800 μg STX equiv. kg^{-1}
ASP (Amnesic Shellfish Poison)	20 mg DA equiv. kg^{-1}
DSP (Diarrhetic Shellfish Poison)	160 μg AO equiv. kg^{-1}

Contaminated shellfish pose a risk to human health, leading to significant illness problems. The syndromes of greatest concern in Portugal due to shellfish toxicity are the Paralytic Shellfish Poisoning (PSP), the Amnesic Shellfish Poisoning (ASP), and the Diarrhetic Shellfish Poisoning (DSP). As the name implies, it can cause paralysis, amnesia and diarrhoea. In Portugal, the Portuguese Institute for the Ocean and Atmosphere (IPMA) carries out the official control of bivalve mollusc production areas, to comply with the legal limits stipulated in Table I by the European Commission [2]. In this way, whenever biotoxin concentrations exceed safety limits, the harvest and trade of shellfish are prohibited, resulting in temporary closures of shellfish production areas.

On the public health and consumer side, these safety limits guarantee that contaminated shellfish do not enter the market. Yet, from the local producer's point of view, this reactive response causes economic losses. Although the bivalves eventually depurate the toxins naturally and reach the safety conditions required by the market, the farmers have to fulfil a certain demand for the product at a certain time. Furthermore, with the climate changes an increase of HAB episodes is expected, which will lead to more frequent and prolonged disruptions of product supply. Therefore, it is essential to develop proactive strategies that mitigate the damage caused by the proliferation of harmful phytoplankton and provide advance warnings to farmers.

B. Motivation

In recent years, the popularity and demand for seafood have increased in Portugal, with fresh seafood expected to have a revenue of 208.4 million euros by the end of 2023 and a market volume amount of 9.5m kg by 2028. However, the sustainability of the shellfish business can be compromised by HAB events, which can cause the temporary closure of shellfish production areas. Robust monitoring programmes must be carried out in the several harvesting areas, to safeguard

public health and minimize the economic losses from the local farmers.

This thesis was developed within the scope of the research project "MATISSE: A Machine Learning-Based Forecasting System for Shellfish Safety", funded by the Foundation for Science and Technology (FCT), with the purpose of exploring new techniques to tackle the problem of contamination of bivalve molluscs. Although forecasting methods give relatively good results and provide advanced information on which bivalve production areas will close, they do not allow us to know from a biological point of view which variables have the greatest impact on HABs. In this way, this thesis was intended to use causality methods to infer causal relationships from time series. More specifically, it aims to work with causality to identify the variables whose past values most influence current biotoxin contamination and, based on these relationships, try to predict future contamination. This new approach seeks to better explain the phenomenon of contamination and the potential predictors of contamination.

Identifying the variables that contribute to shellfish contamination and understanding when production areas can be re-opened is crucial. For instance, some authors have already composed studies about environmental variables and biotoxin accumulation in Portugal, with IPMA data. Pereira (2021) [3] through the platform Dynamic Bayesian Network Online (MAESTRO), discovered that *chl-a* and the sea surface temperature (SST) correlate with ASP in the Ria de Aveiro3 (RIAV3) production area and with PSP-producing phytoplankton in RIAV2. Moreover, it was also shown that *chl-a* between consecutive production areas (RIAV2 and RIAV3) reports a high correlation. Patrício et al. (2022) [4] studied time series of DSP toxins, from which they could identify interesting patterns in Portugal. The authors, in addition to identifying that regions in the North have higher levels of toxicity than regions in the South, also identified a seasonality in the data, reflecting peaks in toxins in May and between August and October. It was also verified that assessing correlations between the same species in different areas is more impactful than assessing correlations of different species in the same area. This analysis is supported by the fact that HAB toxins-producing and species accumulation vary among species. Even under the same conditions, some species can accumulate toxins faster than others, which is in accordance with the concept of indicator species. This term refers to the shellfish species that has the highest rate of toxin accumulation for a given production area because accumulates biotoxins at the fastest rates, such as mussels.

In recent decades, several contributions have come from different domains to infer causality and overcome the typical limitations of machine learning systems, such as correlation-based models and lack of interpretability (black boxes). However, due to the complexity of real-world data, developing causal techniques remains a challenging task. Papana (2021) [6] has compiled an extensive survey of methods, grouping them into two categories that permit analyzing the relationship between variables: (i) non-directional connectivity measures: symmetrical methods that aim to quantify how strong the association between variables is, without indicating the direction of the relationship; (ii) directional connectivity measures: inspired by the concept that the effect comes after the cause, these methods seek to determine the direction of the relationship. Roughly speaking, the author compares correlation methods against causality methods, suggesting that causality measures were better than correlation measures. The results emphasised that correlation tends to fail when the data has temporal dependencies.

Causal discovery techniques can play a pivotal role by inferring causal relationships between variables. Understanding causal relationships brings us closer to comprehending and characterizing dynamic systems, allowing us to potentially foresee the effects of environmental system changes before they happen. It is worth mentioning that some of the findings resulted in the following approved article in a Portuguese conference: Ribeiro, D., Ferraz, F., Lopes, M. B., Rodrigues, S., Costa, P. R., Vinga, S. and Carvalho, A. M. (2023). Causal graph discovery for explainable insights on marine biotoxin shellfish contamination. (Submitted and approved (Submitted and approved at *24th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*)).

II. Background

A. Causation and Causal Discovery

Throughout the centuries, mankind has always searched for an explanation about the world and its surroundings. Determining and understanding how a particular event can influence another has motivated many discussions over the years. This curiosity concerning causal relationships has been a study case for numerous academics, from a purely abstract and philosophical term to a concept in statistics and computer sciences.

There is no universal definition. Causality or causation are concepts used for referring to cause-and-effect relationships. The cause is responsible for creating an effect, therefore, the cause is required for the effect to exist, but the opposite is invalid. This analysis enables the exploitation of the causal mechanisms underlying the data-generating processes.

Causality research can be divided into causal discovery and causal inference. The first analyse and describe the causal relationships in the data, and the latter estimates the causal effect of treatment on the outcome, i.e., the potential effects arising when there are changes in the system.

This thesis primarily focuses on causal discovery. Causal discovery, or structure learning, is the task of searching for causal patterns. These models are responsible for analysing, discovering and illustrating the relationships inherent to observational data. In this way, the mechanism behind complex dynamic systems is described through causal interactions. Unlike correlation-based methods, which only use associative relationships, causality algorithms enable the reconstruction of a network with a typology with directions (causal associations). Understanding the information flow is essential in many areas, such as biomedical, climate research or business.

B. Strucutral Causal Models

The structural causal model (SCM) is the common framework employed by causal discovery algorithms to represent

causal relationships intuitively, both from a probabilistic and causal point of view. Let $X = \{X_1, X_2, \cdots, X_d\}$ be a set of n endogenous variables of interest and $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \cdots, \mathcal{U}_n\}$ a set of n exogenous noise variables that represents the causal effect of unobserved confounders (often implicit in graphical models). An SCM, through deterministic equations, describes the causal nature among variables in a DAG, $\mathcal{G}$. Suppose that each endogenous variable X_i can be determined as a causal mechanism/function, f_i, of its parents $\mathcal{P}a_i^{\mathcal{G}} \subseteq X \backslash \{X_i\}$ and exogenous noise $\mathcal{U}_i$ within a causal graph [7]:

$$X_i := f_i(\mathcal{P}a_i^{\mathcal{G}}, \mathcal{U}_i), \qquad i = \{1, \cdots, n\}, \tag{1}$$

where the operator ":=" exhibits the asymmetric essence of the equation to represent the causal dependencies. This set of equations that model an SCM is known as structural equations.

C. Causal Discovery and Time Series Analysis

This thesis focuses on causality for time series data, so no in-depth review of the literature on non-time series will be conducted. Time series can be univariate or multivariate. A univariate time series (UTS), $\{y_t, t = 1, 2, \cdots, T\}$, is a time series with a single variable being measured over time t, while a multivariate time series (MTS), Y_t, is a time series with n UTS measured over time t, which can be denoted as $Y_t = [y_t^1, y_t^2, \cdots, y_t^n]$. The y_t^i represent the i^{th} time-series component at time t.

Causality discovery in time series seeks to discover the causal connections between n-variate time series Y_t and the time lag between a cause and the corresponding effect. Taking into account the addition of the temporal component to the concept of causality, the remaining section will present some of the state-of-the-art methods of causality applied to time series, namely GC, TCDF, PCMCI, VAR-LiNGAM and DYNOTEARS.

a) **Granger Causality:** One of the oldest and most well-known concepts to infer causal relationships from time series data is the Granger causality [8]. This statistical concept verifies whether a time series can be used to improve the prediction about another time series, respecting temporal precedence. Consider two univariate time series *X* and *Y*, *X* Granger causes *Y* if *X* contains unique and statistically significant information on the past observations of *Y* that is not available in *Y*. This can be formalized using a vector autoregressive (VAR) model. It is important to note that this test is bivariate. Therefore, when applied in a multivariate context, it may result in some spurious correlations.

b) **TCDF:** Here, we explore a deep-learning extension of Granger causality: the Temporal Causal Discovery Framework (TCDF) [9] that employs attention-based CNNs to uncover nonlinear causal relationships between observational time series. For a multivariate time series as input, each network predicts a univariate time series. In addition, TCDF models autocorrelation by using the target time series as input. Potential causes are evaluated by the attention mechanism, which filters the most valuable information for prediction by ranking the attention scores from high to low, resulting in a set of potential causes for each time series. A causal interpretation is given after the softmax function σ, followed by a semi-binarization to be applied. The latter selects all attention scores below a threshold s_p. This threshold s_p is achieved by calculating the biggest attention score associated with the largest gap between two adjacent attention scores. If the attention score is below s_p, a time X_i is viewed as a potential cause of the target time series X_j.

A causal validation step is conducted to validate if a potential cause is an actual cause of the predicted time series, distinguishing merely correlation from causation. Potential causes are validated through the permutation importance (PI) test, a permutation-based procedure for identifying significant causal relationships by measuring how much the network loss score increases when the values of a variable are randomly permuted.

Finally, the network kernel weights are analyzed to learn the time delay of established causal relationships. Each input time series is stored in a row, and the importance of each time delay of the associated time series is in a column.

c) **PCMCI:** The PCMCI algorithm [10] is a constraint-based approach that exploits the conditional independencies embedded in the data to build the underlying causal graph that follows several assumptions: causal Markov condition, causal sufficiency, causal stationarity and faithfulness. This causal discovery algorithm is able to detect time-lagged linear and nonlinear causal relations in time-series data and can be framed into two main phases: the Peter-Clark (PC) phase and the Momentary Conditional Independence (MCI) phase. In the first stage, the PC algorithm is applied to uncover dependencies between each variable and estimate a set of parents for every variable. The resulting skeleton is constructed based on conditional independence tests and the unnecessary edges, such as indirect links or independent links, are removed to avoid conditioning on irrelevant variables. At this stage, the goal is to identify direct causal relationships and remove spurious correlations. In the second phase, an MCI test is conducted to determine causal relationships between the previously selected time-shifted parents and each pair of variables by testing for independence, while taking into account autocorrelation. Basically, the direction of causal links is established and the flow of information is identified.

d) **VAR-LiNGAM:** The VAR-LiNGAM [11] is an algorithm that belongs to the noise-based approach, where causal models are described by a set of equations. Each equation describes one variable of the causal structure plus some additional noise.

However, given two variables, it is not easy to distinguish cause from effect. There must be a way to capture the asymmetry between them. The VAR-LiNGAM is a temporal extension assuming linear, non-Gaussian, and acyclic models (LiNGAM), to uncover the underlying asymmetries, i.e., causal relations. Without these additional assumptions, it would not be possible to identify the direction of the causal relationship. The main idea of this algorithm is to estimate the least-squares of the Structural AutoRegressive Model (SVAR) to determine the causal order based on the independence between the residuals and predictors. This step is carried out iteratively by first identifying the predictor that is the most

independent from the residuals of its target variables. Then, the effects of the previously identified predictors are removed.

e) **DYNOTEARS:** The DYNOTEARS algorithm [12] is a score-based approach that aims to learn an SVAR model in time-series data. This model can be considered a class of Dynamic Bayesian Networks (DBNs), which are unable to capture temporal dynamics. DYNOTEARS differs from other existing algorithms by simultaneously learning both contemporaneous (intra-slice) and lagged (inter-slice) dependencies between variables, instead of applying these steps successively. The adjacency matrices for the inter-slice and intra-slice dependencies are learned by minimizing a loss function based on the Frobenius norm of the residuals of a linear model. However, an acyclicity constraint is still necessary to model contemporaneous links.

III. Experimental Analysis

This chapter develops several data simulations employing the causal discovery algorithms introduced in Section II-C and can be divided into four sections. Note that all the algorithms employed were executed using *Python*.

A. Materials Description

Blood-oxygen-level dependent functional magnetic resonance imaging (BOLD FMRI) This real-world benchmark [13] describes the neural activity from different brain regions based on the blood flow change and comprises 28 brain networks. Each brain region receives a time series fed through a nonlinear balloon model, white noise, and hidden external input. Although these data are simulated, they are so rich and complex that they are considered as if they were real in the literature. The experimental analysis will consider only 17 simulations, excluding simulations with 50-time series (50 nodes) and sample sizes greater than 400. These specifications attempt to bring the benchmark data closer to the data from our case study.

CauseEffectPairs Benchmark dataset containing 100 real-world time series. It consists of several bivariate time series to assess who is the cause and effect among 37 domains (e.g. biology, meteorology, medicine, etc.) [14]. For the experimental programme, we select 16 datasets whose properties closely resemble our case study. In other words, datasets with interactions between meteorological and biological variables. Temperature, precipitation and wind speed are some of these features.

B. Setup

From the Granger family, two methods were tested: GC and TCDF. The pairwise Granger causality (GC) used an F-test to compare the full model with the restricted model, considering a significance level (α) of 5%. The TCDF utilizes a convolutional neural network (CNN) with only one layer (no hidden layers), kernel size $K = \tau_{\max} + 1$ and dilation coefficient $c = K$. The tuning parameters to be defined are the training epochs $\in [1000, 2000, 3000, 4000]$, learning rate $\lambda \in [0.001, 0.005, 0.01, 0.05, 0.1, 0.5]$, optimizer $\in$ [Adam, RMSprop] and significance $s \in [0, 0.2, 0.4, 0.6, 0.8, 1]$. From the constraint-based family, two algorithms proposed by the same author were analyzed: PCMCI and PCMCI+. A nomenclature was created for the different settings to facilitate the representation of these methods in analytical formats. 1 and 3 denote the conditional independence test used, namely the Partial Correlation conditional independence test (ParCorr) and Gaussian Processes and Distance Correlation conditional independence test (GPDC), respectively. The "P" from "plus" (+) denotes the algorithm PCMCI+, instead of PCMCI. Finally, the term "C" comes from corrected p-values, which is a setting available in both PCMCI and PCMCI+ to control the false discovery rate. The correction of p-values is available in the *Python* package called *Tigramite* using the function *get_corrected_pvalues(fdr_method="fdr_bh")*. The significance threshold (α_{PC}) must belong to [0.01,0.05,0.1,0.2,0.3,0.4,0.5].

For the noise-based family, the VAR-LiNGAM was performed. The optimal automatic selection of lags was set to none (*criterion=None*), guaranteeing the same maximal time delay for all methods. A prune option is also available to select non-zero adjacencies, with an adaptive LASSO regularization through the BIC criterion (*sklearn.LassoLarsIC(criterion="bic")*). Sometimes, even with a non-zero adjacency matrix, the weight given to certain causal links is minimal. Therefore, the parameter α_{limit} was created to select only causal links with a weight more significant than the α_{limit}. It establishes a sparser graph with causal relations that are more likely to be well-identified.

Finally, from the score-based family, the DYNOTERS algorithm. The causal graph is built based on three parameters. The first, $\lambda_w \in [0.01, 0.1]$ is responsible for applying the L1 regularisation on intra-slice edges. The second, $\lambda_a \in [0.1, 1, 10]$ is responsible for applying the L1 regularisation on inter-slice edges. The third, $w_\tau \in [0.01, 0.1]$, employs a fixed threshold for absolute edge weights.

For this analysis, we consider a $\tau_{\max} = 5$ for all methods.

C. Evaluation Measures

Validation of causality algorithms is essential to evaluate a time series causal graph. This thesis will consider a binary classification to evaluate the F1 score between adjacent nodes. Thus, the causal links between the true underlying structure (actual class) given by the benchmark datasets and the graphs resulting from the algorithms (predicted class) are compared with the following elements:

- True Positive (TP): Causal link that the algorithm predicted and belongs to ground truth.
- False Positive (FP): Causal link that was predicted but does not belong to the underlying structure.
- True Negative (TN): Causal link that was not predicted and does not belong to ground truth.
- False Negative (FN): Causal link that was not predicted but should belong to the proper underlying structure.

Considering the previous notation (TP, FP, TN, FP) for binary classification, the precision, recall, F1-score (F1) and

false positive rate (FPR) are obtained:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{4}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TP} + \text{FP}}, \tag{5}$$

D. Experimental Results

This section demonstrates the steps followed to decide which method is most suitable for our case study. One of the biggest challenges when choosing a particular method is which hyperparameters to select. To this end, a grid search was conducted with datasets with different particularities and degrees of complexity to determine which hyperparameters were used most frequently, valuing their ability to adjust to real data. Note that the models that presented phase in this limited F1 score for their better parameterization were discarded, namely GC, 1, 1C, 1PC, 3, 3P and 3PC. Thus, The best hyperparameters for each model are $\alpha_{\text{PC}} = 0.3$ for the 1P and 3P, $\alpha_{\text{limit}} = 0.1$ for VARLINGAM, $w_\tau = 0.3$, $\lambda_w = 0.01$, $\lambda_w = 1$ for DYNOTEARS and epochs= 2000, $\lambda = 0.5$, $s = 1$ and Adam as optimizer for TCDF. Table II presents the results obtained using the previously mentioned parameters. The best F1-score was 0.495, obtained by PCMCI+ 3P with $\alpha_{\text{PC}} = 0.3$ and will be the model used in the case study.

TABLE II: Selection of the algorithm for the case study. Bold represents the highest F1-score.

	1P	3P	DYN	VL	TCDF
1	0.29	0.22	0.44	0.50	0.90
2	0.26	0.17	0.53	0.18	0.59
3	0.39	0.32	0.43	0.20	0.59
8	0.40	0.67	0.31	0.13	0.70
10	0.22	0.40	0.53	0.38	0.82
11	0.12	0.11	0.47	0.13	0.53
12	0.33	0.29	0.47	0.13	0.67
13	0.31	0.57	0.48	0.32	0.57
14	0.22	0.00	0.40	0.35	0.80
15	0.29	0.44	0.63	0.40	0.57
16	0.43	0.31	0.70	0.35	0.75
17	0.11	0.32	0.50	0.18	0.70
18	0.67	0.75	0.50	0.40	0.50
21	0.29	0.22	0.44	0.50	0.90
22	0.00	0.22	0.31	0.25	0.63
23	0.31	0.55	0.50	0.43	0.45
24	0.73	0.36	0.40	0.35	0.50
0001	1.00	1.00	1.00	1.00	0.00
0002	0.00	1.00	1.00	1.00	0.00
0003	0.67	0.67	1.00	1.00	0.00
0004	0.00	0.67	1.00	1.00	1.00
0020	0.40	0.00	0.00	1.00	0.00
0021	0.00	0.00	0.00	0.00	0.00
0048	0.67	1.00	0.00	1.00	0.50
0049	0.67	1.00	0.00	0.00	0.40
0050	1.00	0.67	0.00	0.00	0.40
0051	1.00	0.00	0.00	0.00	0.00
0053	0.31	0.40	0.22	0.57	0.40
0077	1.00	1.00	0.00	1.00	0.50
0081	0.67	0.67	0.00	0.00	0.50
0082	1.00	1.00	0.00	0.00	0.50
0083	1.00	0.67	0.00	0.00	0.40
0093	1.00	0.67	0.00	0.00	0.00
Final Mean	0.476	**0.495**	0.371	0.387	0.478

IV. Case Study

This Chapter is divided into two sections. The first concerns the data preprocessing and the selection of production areas and species to be engaged for the causal analysis. The second will discuss the results obtained for the causal investigation of DSP toxins in mussels, donax clams and cockles.

A. Data Preprocessing

This section presents a framework for collecting, processing and analyzing multivariate time series corresponding to each production area. This process is conducted through five stages: data acquisition and integration, data cleaning, data selection, data imputation and data transformation.

1) Data Acquisition and Data Integration: The Portuguese coast has 41 production areas, each containing a multivariate time series that describes the temporal behaviour of several biological and meteorological features. Each multivariate time series was obtained by collecting multiple time series from IPMA and Copernicus, and an integration process allowed the aggregation of numerous multivariate series.

The time series were collected from 2015 to 2020 from two sources, namely IPMA and Copernicus. Note that the variables gathered were not chosen randomly. Several environmental variables have been pointed as potential predictors of biotoxin shellfish contamination, such as atmospheric temperature, sea surface temperature (SST), rainfall, phytoplankton and wind direction.

IPMA *In-Situ* and Meteorological Data IPMA provides both *in-situ* and meteorological measurements. The weekly *in-situ* data concedes the concentration of biotoxins in different species of bivalve molluscs (ASP, DSP and PSP toxins) and phytoplankton cell counts (ASP, DSP and PSP toxins-producing phytoplankton) from each shellfish production area. The sampling sites are strategically placed within shellfish production areas in recognized phytoplankton accumulation zones. As species of bivalve molluscs accumulate toxins with different accumulation rates, for the same production area, we can have, for example, two sampling locations, one for mussels and another for cockles. The daily meteorological data is collected from 22 meteorological stations, measuring the minimum, mean, and maximum air temperature, mean wind intensity, mean wind direction, wind direction (encoded in cardinal directions) and rainfall. For each production area, the nearest meteorological station is assigned using the station coordinates and the coordinates of the sampling points.

Copernicus Remote Sensing Data Copernicus, previously known as GMES (Global Monitoring for Environment and Security), is the Earth observation component of the European Union's space programme that delivers accurate data to help understand the consequences of climate change and enhance environmental management [15]. In particular, the Copernicus Marine Environment Monitoring Service (CMEMS) provides diary *chl-a* concentration and SST measurements by remote

sensing (satellite data). Regarding the acquisition process, this satellite data has several uniformly spaced grid points that allow determining which is closest to the studied sampling location. The grid points closest for *chl-a* and SST are only accepted if they have a small or no number of missing data.

Time Series Integration The sampling times must be considered to integrate the time series data from IPMA and Copernicus. The meteorological and satellite data were measured approximately daily, while the *in-situ* data were approximately weekly. However, there were weeks without any measurements and weeks with multiple sizes. Thus, the data were resampled over 312 weeks, capturing the worst situation of toxins and toxic phytoplankton each week and the weekly mean for the other variables. Weeks without any observations were considered missing data.

2) Data Cleaning: Initiating with the biotoxins time series, it contained some categorical codes assigned by IPMA, represented by ND (non-detected), NQ (non-quantifiable) and NR (not-done). The first two codes refer to levels of toxins concentration below the minimum required concentration for the methodologies employed. Depending on the type of biotoxin measured, the ND and NQ codes were replaced by 71 μg STX equiv. kg^{-1} in the case of PSP toxins, 1.8 μg DA equiv. kg^{-1} for ASP toxins and 36 μg AO equiv. kg^{-1} for DSP toxins. The toxins assigned with the NR code were treated as missing values since the samples were not analyzed. Furthermore, the data contains manual input errors, such as the name of bivalve species or production areas, which must be corrected. errors. In the total data, 45 sample points were misspelled.

Concerning the phytoplankton time series, IPMA assigned a <LD code or null value to samples with cell counts lower than the minimum concentration required for the detection methodologies. These codes were replaced by 20 cell L^{-1} (fixed threshold given by IMPA) for all phytoplankton species.

For the meteorological time series, IPMA assigned the code -990 to represent observations without measurements. Thus, the observations were replaced by missing data for later imputation.

3) Data Selection: At this stage, the data includes 41 multivariate time series corresponding to each production area. As contamination is a spontaneous phenomenon with no periodicity, depending on the type of toxin that produces the contamination and the environmental variables, the bivalve mollusc harvesting areas are affected differently. Causal discovery algorithms may have the ability to infer causal relationships from a set of observational data, but the presence or absence of a high level of contamination alters the intrinsic relationship between these variables. For instance, if in a given production area, the ASP toxin only registered three contamination cases, the causal algorithm would only represent the common patterns and ignore these isolated regimes. To conduct a causal analysis of HAB events, exploring production areas with high contamination levels is essential. Thus, the general behaviour of these algorithms would be to determine the drivers of most toxicity in that region. Therefore, only the DSP toxins make sense to analyse due to the low number of toxic events on PSP and ASP toxins. Since toxins accumulation rates vary among HAB toxins-producing and bivalve species, each production area should be evaluated individually. This means the values obtained must be segregated by area and species. Figure 1 represents the cases of DSP contamination with the most events per species.

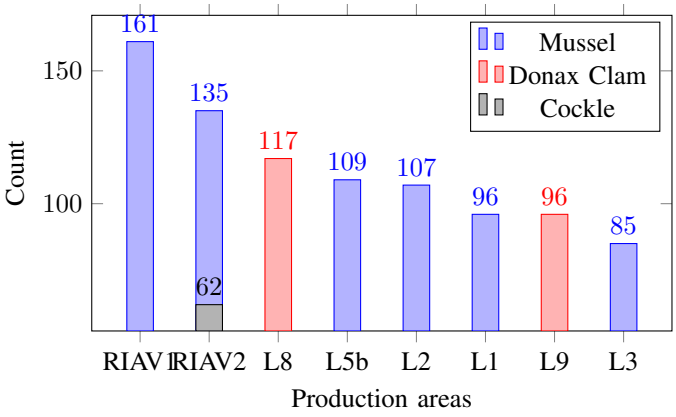


Fig. 1: Highest DSP concentration events above the safety limit per production area.

Once the most promising areas have been chosen, it remains to be seen whether the amount of missing data could be a limitation to the variables that have been provided. The features associated with the ASP and PSP toxins have already been removed, and the wind direction in discretised format was removed because the independence test of the 3P algorithm (GPDC) only accepts continuous data. The minimal and mean air temperatures were removed due to temporal similarity with maximal air temperature, which can lead the causal algorithm to some mistakes. Concerning the variables in the Table, for a variable to be accepted, it must contain a maximum of 35% missing data (approximately 1/3 of the data). The variables shown in bold have been removed from the respective time series. In the case of production area L3, as the DSP toxin, which is the target variable, has a lot of missing data (62%>35%), L3 was removed from this set of selected production areas.

4) Data Imputation: A semi-automatic algorithm was implemented to select the best imputation method for the missing values. First, the longest period without missing data was calculated. The corresponding values were considered as the ground truth (or true values y). Then, random missing data was inserted into the calculated period. Finally, the interpolate function from the *pandas* library in *Python* was used with the following options: *linear*, *quadratic*, *cubic*, *spline*(order=2), *spline*(order=3), *locf*, *nocb*, *nn*, and *knn*. The imputed values are seen as the predictor, and the best method is the one that minimizes the root mean squared error (RMSE).

Steps 2 and 3 were performed for three different levels of complexity: 10%, 20% and 30% of missing data. Thus, the method that best minimizes this error is selected. Note that this procedure must be applied to each variable.

5) Data Transformation: Two data transformations were performed: data stationary and data normalization.

Data Stationarity A stationary analysis was conducted to avoid the spurious correlations that non-stationary time series may introduce.

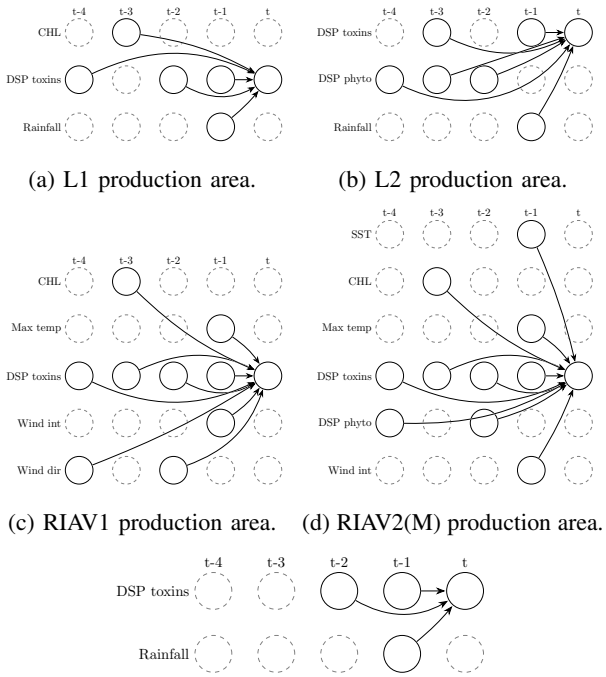
For analysing stationarity, a combination of the augmented Dickey-Fuller (ADF) test [16] test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [17] was performed. This system can produce one of four results: (i) the time series is stationary; (ii) the time series is trend stationary: detrend by applying a moving average; (iii) the time series has a stochastic trend: difference one time; (iv) the time series is not stationary: difference n times (other transformations may be necessary). In practice, only the first three were found, and when necessary, we performed detrending (ii) or differencing (iii) to make the time series stationary. These techniques are available separately in the *Python* package called *statsmodels*, using the *adfuller* and *kpss* functions, for the ADF and KPSS tests, respectively.

Data Normalization Finally, normalization techniques were performed using the *Python* package called *sklearn*, using the function *minmaxScaler* so that variables with different scales could contribute equally to the model.

B. Results and Discussion

1) Mussel Analysis: The mussels *Mytilus spp.* with the most cases of DSP toxins contamination are in the following production areas: L1, L2, RIAV1, RIAV2(M) and L5b. After carrying out the data preprocessing steps listed in Section IV-A, the causal algorithm selected in Section III was used to illustrate in Figure 2 the causal relationships detected between contaminated mussels and environmental variables, up to a time lag of 4 weeks. The RIAV2(M) production zone has the most contaminating variables, followed by RIAV1 and L2. It should be noted that RIAV1 is the production area that most often exceeds safety limits, which may explain the difficulty in finding a causal link with phytoplankton. Additionally, the quantification of toxic phytoplankton may not be representative of the entire water column where mussels inhabit, other DSP-toxins producing algae may be present, and more plausible is the differential toxins accumulation dynamics of shellfish that may not mirror algae abundance, as have been observed in controlled experiments in the laboratory.

This approach alone does not allow us to isolate which predictors contribute most to mussel contamination above safety limits. Still, it does allow us to more generally infer which are the most important variables in the toxicity ecosystem and which temporal lag impacts mussel toxicity. By selecting production areas with more contamination cases, we are closer to achieving characteristics related to high toxicity. Thus, Figure 3 proposes two ways to aggregate the results from mussel contamination: one that values whether a given causal link is found in several production areas with the same time lag (Figure 3b) and another that, instead of penalizing contaminants detected with different temporal lags, it benefits if the same predictors exist in different production areas (Figure 3a). Without discretizing the temporal lags, we can state that the DSP autocorrelation presents good evidence of being able to justify its current toxicity values based on past values. Furthermore, DSP toxins-producing phytoplankton and rainfall are the best options to complement the analysis. Discretizing the temporal lags, we can conclude that the maximum



(a) L1 production area. (b) L2 production area.

(c) RIAV1 production area. (d) RIAV2(M) production area.

(e) L5b production area.

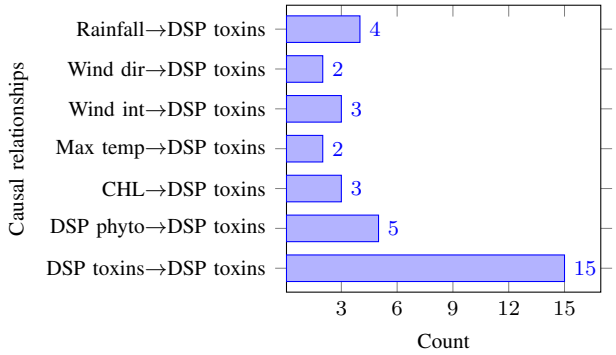
Fig. 2: Causal relationships detected in contaminated mussels with a temporal lag up to four weeks.

temperature, wind intensity and rainfall are predictors of toxin concentration for shorter-term dependencies (1-week and 2-week lags) and *chl-a* for longer-term dependencies (3-week lag). Although DSP toxins autocorrelation is found with the four timestamps, only a 1-week and 2-week lag is revealed in 4 or 5 areas. Thus, it has a greater incidence in relationships of shorter-term dependencies. Note that the results with discretized temporal lags are obtained from Figure 2.

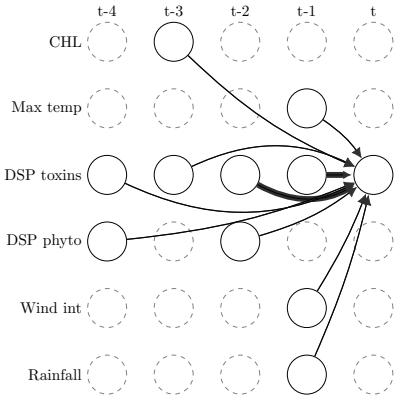
Since mussels are the indicator specie in most production areas, more samples are available, so an estuarine-lagoon and coastal analysis is conducted. Further, given their coastal characteristics, production areas L1, L2 and L5b are considered coastal zones, and RIAV1 and RIAV2(M) are considered estuary-lagoon zones. Figure 4b presents the aggregated results for the estuarine-lagoon zones. Only short-term connections were detected, namely rainfall and the autocorrelation of DSP toxins with a 1-time lag. DSP toxins-producing phytoplankton report the same importance as rainfall in the approach without discretized temporal lags (Figure 4a). The fact that phytoplankton does not appear in the discretized results only means that it was detected with three different lags in Figure 2.

Figure 4d indicates that the DSP toxins present a very strong autocorrelation for coastal areas, highlighting 1-week and 2-week. Additionally, it was discovered that *chl-a*, maximum air temperature and wind intensity do not count as contaminants for coastal areas. Furthermore, rainfall does not count as a predictor of contamination for estuarine-lagoon areas.

Mussel Summer/Winter Analysis: Finally, an analysis of contamination during summer and winter is carried out. The objective would be to understand if there is any hidden regime

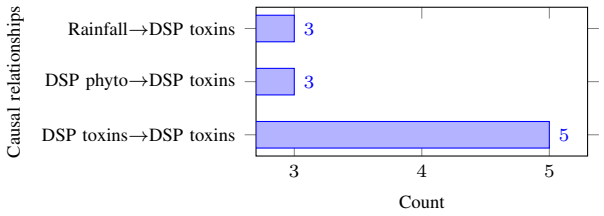


(a) Without discretized temporal lags.

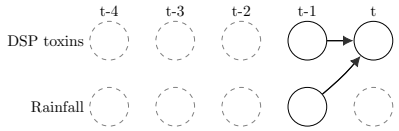


(b) With discretized temporal lags. There are thicker edges to record whether causal links were found in 4 or 5 areas and thinner ones for 2 or 3.

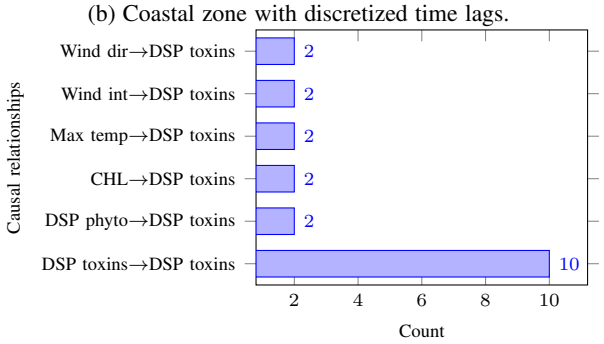
Fig. 3: Aggregated results for mussels contamination with discretized lags (a) and without discretized lags (b).



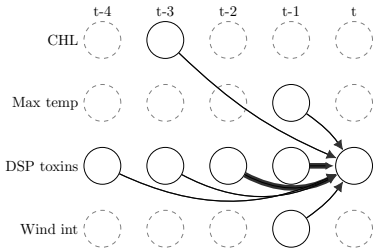
(a) Coastal zone without discretized temporal lags



(b) Coastal zone with discretized time lags.



(c) Estuarine-lagoon zones without discretized temporal lags.



(d) Estuarine-lagoon zones with discretized temporal lags. There are thicker edges to record whether causal links were found in 4 or 5 areas and thinner ones for 2 or 3.

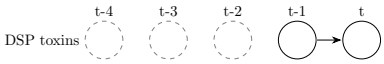
Fig. 4: Aggregated results for coastal and estuarine-lagoon zones in mussels contamination with discretized lags (a) and without discretized lags (b).

in the data during this period. To this end, the summer period was considered between 20/03 - 23/09, and the winter period between 01/01 - 20/03 and 23/09 - 31/12. A comparison will be conducted against Figure 3b, which illustrates the six years of contamination.

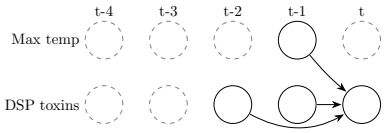
In the summer period, only the wind direction with a 4-week time lag is added. On the other hand, *chl-a*, maximum air temperature and DSP toxins-producing phytoplankton are no longer detected. Wind intensity, rainfall and DSP toxins continue to show causal relationships, but with changes in the timestamp at which they are detected. To summarize, rainfall and DSP toxins autocorrelation stand out due to short-term dependencies and wind direction and intensity as long-term relationships. Concerning the winter period, *chl-a*, DSP toxin-producing phytoplankton, wind intensity and rainfall are no longer detected. Maximum air temperature and DSP autocorrelation manifest mostly as short-dependency.

2) Donax Clam Analysis: The donax clam *Donax trunculus* with the most cases of DSP toxins contamination are in the following production areas: L8 and L9. Figure 5 shows the causal relationships detected between contaminated donax clams and environmental variables, up to a time lag of 4 weeks. The L9 production zone has the most contaminating variables, with the maximum air temperature with a 1-week temporal lag. The lack of causal links in donax clams may be fundamentally due to two factors: (i) geographical position: the L8 and L9 zones are located in the South of Portugal, and, therefore, the same contaminants that were tested in the North may have characteristics many different. In addition to the different ecosystem characteristics, salinity, pH of the water, etc; (b) Toxicity accumulation rate: different species of molluscs bivalves have different sensitivities to contamination.

Donax Clam Summer/Winter analysis Regarding the analysis of donax clam contamination during summer and winter, only 1-week autocorrelation of DSP toxins was detected. This does not mean that there is no intrinsic regime for each production zone, but it does mean that there is no common



(a) L8 production area.



(b) L9 production area.

Fig. 5: Causal relationships detected in contaminated donax clam with a temporal lag up to four weeks.

regime between the two production areas. It would also not be expected that many contaminants would be found, since in Figure 5, only one DSP toxin contaminant was found (in addition to DSP toxins autocorrelation).

3) Cockle Analysis: The cockle *Cerastoderma edule* with the highest number of cases of DSP toxins contamination is in the RIAV2(B) production area. This analysis will be the most restricted because it only concerns one production area, even in terms of validating results with adjacent areas. Figure 6 pictures the causal relationships detected between contaminated cockles and environmental variables, up to a time lag of 4 weeks. Biological variables, namely DSP toxins-producing phytoplankton and DSP toxins, are more active in cockle contamination. Besides, the maximum air temperature manifests as a causal link with short dependence, while *chl-a* presents as a causal link with long causal dependence.

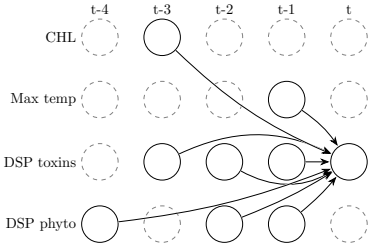


Fig. 6: Causal relationships detected in contaminated cockles in the RIAV2(B) production area with a temporal lag up to four weeks.

Cockle Summer/Winter analysis To conclude, for both summer and winter, the maximum air temperature was lost as a driver of contamination. Additionally, DSP toxins only show autocorrelation up to 2 weeks, being more incisive in summer than in winter. In the summer period, the *chl-a* stands out as a longer-term dependency, while DSP toxins-producing phytoplankton appears as a short-dependent contaminant. On the contrary, in winter, DSP toxins-producing phytoplankton presents a causal link with longer-term dependency.

V. Conclusion and Future Work

Bivalve mollusc production provides healthy food and essential economic support for remote families and small businesses in rural areas. The sustainability of the shellfish business can be compromised by HAB events, which produce high concentrations of biotoxins that can accumulate and contaminate shellfish, making them unsafe for human consumption and leading to significant illness problems. Whenever biotoxin concentrations exceed safety limits, the harvest and trade of shellfish are prohibited, resulting in temporary closures of shellfish production areas.

Understanding causal relationships brings us closer to comprehending and characterising dynamic systems, allowing us to potentially foresee the effects of environmental system changes before they happen. Since shellfish contamination is a real-world problem without knowledge about the ground truth, benchmark data will be used to validate several state-of-the-art methods, including GC, TCDF, PCMCI, VAR-LiNGAM and DYNOTEARS. Two benchmarks were employed: the BOLD FMRI and CauseEffectPairs, composing 47 times series data. A grid search was conducted for each time series to obtain the best hyperparameters for each model based on the F1-score. Ultimately, the PCMCI+ parameterization using GPDC with $\alpha_{\text{PC}} = 0.3$ stood out.

Regarding mussel *Mytilus spp.* contamination, the RIAV2(M) production area has the most contaminants, followed by RIAV1 and L2. Discretising the time lags, we can conclude that maximum temperature, wind intensity, rainfall and the autocorrelation of DSP toxins are predictors of the concentration of DSP toxins for shorter-term dependencies (lags of 1-week and 2-weeks) and *chl-a* for longer-term dependencies (lag of 3-weeks). In addition, the results were aggregated by coastal and estuarine-lagoon zones, where it was found that apart from the 1-week autocorrelation of DSP toxins, only rainfall impacts coastal zones 1-week in advance. In contrast, estuarine-lagoon zones show *chl-a*, maximum air temperature and wind intensity as potential predictors of contamination, and DSP toxins with a very strong autocorrelation, being present up to a 4-week time lag. Finally, an analysis of mussel contamination in winter and summer was carried out to determine hidden regimes in the data. In summer, rainfall and DSP toxins autocorrelation stand out due to short-term dependencies and wind direction and intensity as long-term relationships. In winter, maximum air temperature and DSP toxins autocorrelation manifest primarily as short dependency causal links.

Concerning donax clams *Donax trunculus* contamination, only two potential contaminants were inferred at short notice: the 1-week maximum air temperature and the DSP toxins. The lack of causal links is likely due to the South position of the L8 and L9 production areas or to the slower accumulation rate of the donax clams, making it more difficult to infer these associations.

Finally, the cockle *Cerastoderma edule* is strongly connected with biological variables, namely DSP toxins-producing phytoplankton and DSP toxins. Additionally, a maximum air temperature manifests with a temporal lag of 1-week and chl-a with a temporal lag of 3-weeks. Interestingly, in summer, the DSP toxins-producing phytoplankton plays a more immediate role in toxicity, while in winter, they only become noticeable 4-weeks in advance.

In conclusion, causality methods aim to provide an interpretability component that prediction algorithms cannot provide. Time series causality discovery applied to a real-world context, such as the contamination of bivalve molluscs, aims to identify the potential environmental variables with the most significant influence on the concentration of biotoxins in shellfish. In the future, increasing the number of production areas to be studied would be very interesting. With more production areas, a more concise analysis of the main drivers of contamination by coastal zone and estuarine-lagoon zone could be carried out, in addition to an exploration by geographic zone. Indeed, potential predictors of contamination may present different characteristics if studied further north of the country or further south. Moreover, with more production areas, a more concise approach to contamination by bivalve species could be carried out. Mussels can be an indicator species in many production areas but sometimes negatively discriminate against other species due to long contamination periods. Finally, it would be beneficial to study the spread of toxins between adjacent production zones.

ACKNOWLEDGMENT

I want to thank my supervisors, Alexandra Carvalho and Susana Vinga, for all their support, guidance and availability over the last year, and everyone who made themselves available and contributed to this project. I want also to thank my family and friends for their support, motivation and, above all, patience during this life journey. Finally, I would also like to acknowledge the Foundation for Science and Technology (FCT) funding through the projects UIDB/50008/2020 (Institute of Telecommunications), UIDB/50021/2020 (INESC-ID), and the project "MATISSE: A Machine Learning-Based Forecasting System for Shellfish Safety" (DSAIPA/DS/0026/2019). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951970 (OLISSIPO project).

REFERENCES

[1] Mateus, M., Fernandes, J., Revilla, M., Ferrer, L., Villarreal, M., Miller, P., Schmidt, W., Maguire, J., Silva, A. & Pinto, L. Early Warning Systems for Shellfish Safety: The Pivotal Role of Computational Science. *Computational Science – ICCS 2019*. pp. 361-375 (2019)

[2] European Parliament, C. Regulation (EC) No 853/2004 of the European Parliament and of the Council of 29 April 2004. *Official Journal Of The European Union*., https://www.ipma.pt/pt/bivalves/docs/index.jsp

[3] Silva Pereira, A. Time Series Analysis and Forecasting of Shellfish Contamination and Safety. (IST University of Lisbon,2021,10)

[4] Patrício, A., Lopes, M., Costa, P., Costa, R., Henriques, R. & Vinga, S. Time-Lagged Correlation Analysis of Shellfish Toxicity Reveals Predictive Links to Adjacent Areas, Species, and Environmental Conditions. *Toxins*. **14** (2022), https://www.mdpi.com/2072-6651/14/10/679

[5] Vale, P., Gomes, S., Botelho, M. & Rodrigues, S. Monitorizaçao de PSP na costa portuguesa através de espécies-indicadoras. *Avances Y Tendencias En Fitoplancton Tóxico Y Biotoxinas, 2008-01-01, ISBN 978-84-96997-06-6, Pags. 177-180*. (2007,1)

[6] Papana, A. Connectivity Analysis for Multivariate Time Series: Correlation vs. Causality. *Entropy*. **23** (2021), https://www.mdpi.com/1099-4300/23/12/1570

[7] Guo, R., Cheng, L., Li, J., Hahn, P. & Liu, H. A Survey of Learning Causality with Data: Problems and Methods. *ACM Comput. Surv.*. **53** (2020,7), https://doi.org/10.1145/3397269

[8] Granger, C. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*. **37**, 424-438 (1969), http://www.jstor.org/stable/1912791

[9] Nauta, M., Bucur, D. & Seifert, C. Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning And Knowledge Extraction*. **1**, 312-340 (2019), https://www.mdpi.com/2504-4990/1/1/19

[10] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Dino Sejdinovic Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*. **5**, eaau4996 (2019), https://www.science.org/doi/abs/10.1126/sciadv.aau4996

[11] Hyvärinen, A., Zhang, K., Shimizu, S. & Hoyer, P. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal Of Machine Learning Research*. **11**, 1709-1731 (2010), http://jmlr.org/papers/v11/hyvarinen10a.html

[12] Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P. & Aragam, B. DYNOTEARS: Structure Learning from Time-Series Data. *Proceedings Of The Twenty Third International Conference On Artificial Intelligence And Statistics*. **108** pp. 1595-1605 (2020,8,26), https://proceedings.mlr.press/v108/pamfil20a.html

[13] Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J. & Woolrich, M. Network modelling methods for FMRI. *NeuroImage*. **54**, 875-891 (2011), https://www.sciencedirect.com/science/article/pii/S1053811910011602

[14] Mooij, J., Peters, J., Janzing, D., Zscheischler, J. & Schölkopf, B. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal Of Machine Learning Research*. **17**, 1-102 (2016), http://jmlr.org/papers/v17/14-518.html

[15] Copernicus: https://www.copernicus.eu/

[16] Dickey, D. & Fuller, W. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *JASA. Journal Of The American Statistical Association*. **74** (1979,6)

[17] Kwiatkowski, D., Phillips, P., Schmidt, P. & Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal Of Econometrics*. **54**, 159-178 (1992),