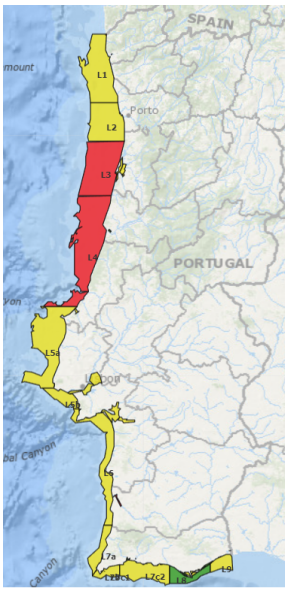
TÉCNICO LISBOA



Causal graph discovery for explainable insights on marine biotoxin shellfish contamination

Diogo Rafael Esteves Ribeiro

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Alexandra Sofia Martins de Carvalho
Prof. Susana de Almeida Mendes Vinga Martins

Examination Committee

Chairperson: Prof. Pedro Filipe Zeferino Aidos Tomás
Supervisor: Prof. Alexandra Sofia Martins de Carvalho
Member of the Committee: Prof. Marta Isabel Belchior Lopes

30 Nov 2023

Declaration

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 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).

Resumo

A proliferação de algas nocivas é um fenómeno natural que provoca a contaminação de bivalves moluscos devido à rápida acumulação de biotoxinas marinhas. O Instituto Português do Mar e da Atmosfera monitoriza regularmente o fitoplâncton tóxico e encerra temporariamente as zonas de produção de marisco sempre que a concentração de biotoxinas excede os limites de segurança. As técnicas de causalidade aplicadas a séries temporais multivariadas permitem identificar as variáveis que mais influenciam a contaminação. Desta forma, vários métodos foram validados em conjuntos de dados de referência, destacando-se o algoritmo de Independência Condicional Momentânea de Peter-Clark +. Com base neste algoritmo, foi explorada a concentração de biotoxinas em mexilhões, conquilhas e berbigões em dados ambientais do IPMA e do Copernicus. Concluímos que a temperatura máxima, a intensidade do vento e a precipitação são factores de previsão da contaminação do mexilhão com toxinas DSP para dependências de curto prazo e *chl-a* de longo prazo. Relativamente à contaminação da amêijoa, apenas foi inferida a temperatura máxima do ar com um desfasamento temporal de uma semana. A contaminação do berbigão mostrou uma grande relação com as variáveis biológicas, nomeadamente o fitoplâncton produtor de toxinas DSP e as toxinas DSP. Adicionalmente, a temperatura máxima do ar manifesta-se com um desfasamento temporal de uma semana e a *chl-a* com um desfasamento temporal de três semanas. Este estudo propõe uma nova abordagem para inferir relações entre variáveis ambientais, melhorando a tomada de decisão e a segurança da saúde pública relativamente ao consumo de marisco em Portugal.

Palavras-chave: Séries Temporais Multivariadas, Contaminação de Bivalves, Biotoxinas Marinhas, Saúde Pública, Modelos Causais, Descoberta Causal

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.

Keywords: Multivariate Time Series, Shellfish Contamination, Marine Biotoxins, Public Health, Causal Models, Causal Discovery

Contents

List of Tables

List of Figures

Nomenclature

ASP Amnesic Shellfish Poisoning

BOLD FMRI Blood-oxygen-level dependent functional magnetic resonance imaging

chl-a chlorophyll-a

CMEMS Copernicus Marine Environment Monitoring Service

CNN Convolutional Neural Network

DAG Directed Acyclic Graph

DBN Dynamic Bayesian networks

DSP Diarrhetic Shellfish Poisoning

DYNOTEARS Dynamic Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning

FFNN Feed-Forward Neural Network

FN False Negative

FP False Positive

FPR False Postive Rate

GC Granger Causality

GPDC Gaussian Processes and Distance Correlation conditional independence test

HABs Harmful Algal Blooms

IPMA Portuguese Institute for the Ocean and Atmosphere

LSTM Long Short-Term Memory

MAESTRO Dynamic Bayesian Network Online

MATISSE Machine Learning-Based Forecasting System for Shellfish Safety

MTS Multivariate Time Series

MVGC Multivariate Granger Causality

ParCorr Partial Correlation conditional independence test

PCMCI Peter-Clark Momentary Conditional Independence

PCMCI Peter-Clark

PSP Paralytic Shellfish Poisoning

SCM Structural Causal Model

SST Sea Surface Temperature

SVAR Structural Vector Autoregressive

TCDF Temporal Causal Discovery Framewor

TN True Negative

TP True Positive

UTS Univariate Time Series

VAR Vector Autoregressive

VAR-LiNGAM Vector Autoregressive Linear Non-Gaussian Acyclic Model

Chapter 1

Introduction

1.1 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, 2].

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 [3]. 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 [4–6].

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 [7].

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 1.1 by the European Commission [8]. 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.

1.2 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 [1]. 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.

Furthermore, several countries adopted HABs forecast-based monitoring programmes to develop sustainable shellfish aquaculture [6, 9, 10]. For instance, in Ireland, a weekly HAB bulletin has been published since 2013, providing short-term forecasts (3-5 days) on the potential development of HAB events, from hydrodynamic models and satellite data. In Scotland, the bulletins have been produced since 2014 and the predictions on the biotoxin concentrations are said to be 74% accurate [11]. The public software is implemented with a "traffic light" based system, with a green colour to indicate that all HAB species and associated shellfish biotoxins are below legal limits, yellow to exhibit that at least one shellfish biotoxin or HAB species presents high toxicity (still below the safety limit) and red suggesting that at least one specie is exceeding the toxicity level threshold. Finally, Portugal employed a warning system to predict HAB and shellfish closure areas, also using the "traffic light" system for better interpretation. This program, in addition to illustrating the current toxicity condition in each shellfish production area based on *in situ* and satellite data, also forecasts the condition of shellfish production areas based on phytoplankton cells, biotoxins concentration, ocean circulation and temperature [12].

Thus, trustworthy predictive information enables the ability to predict the occurrence of toxic events in advance, acting as a warning system for HABs formation [12]. Prior knowledge gives the producers the capability to make better management decisions and reduce human health incidents [9]. Therefore, forecasting models play a fundamental role in predicting these toxic events.

Lee et al. (2003) [13] trained with one hidden layer a Feed-Forward Neural Network (FFNN) to predict one-week ahead algal blooms in Hong Kong. Multiple tests were conducted with different inputs, including chlorophyll-a (*chl-a*), water temperature, solar radiation, rainfall and wind speed. The best scores were achieved using only the time-lagged algal as input.

Table 1.1: Limits regulated by the European Commission. Adapted from [8].

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}

Yussof et al. (2021) [14] trained a long short-term memory (LSTM) network and a Convolutional Neural Network (CNN) with satellite data and *chl-a* concentration from 2003 to 2018, in the West Coast of Sabah, Malaysia. The results showed that the CNN was outperformed by the LSTM, to forecast HABs eight days in advance.

Cruz et al. (2021) [15] conducted a comprehensive literature review of the latest forecasting methods to predict HABs and shellfish contamination. Based on that information, for the Portuguese coast scenario, Cruz et al. (2022) [16], focused on predicting DSP toxins concentrations in mussels one to four weeks in advance with multiple forecasting models. The most accurate response was achieved using a bivariate LSTM network predicting biotoxin contamination 1-week in advance based on toxic phytoplankton and biotoxin concentration time-series data.

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) [17] 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. Madeira (2022) [18] used a joint analysis between cross-correlation and DBNs to analyze pairs of biotoxin concentration or pairs of phytoplankton cell counts between adjacent shellfish production areas in southern Portugal. The idea is to determine whether, after a given area reaches high levels of contamination, the adjacent area follows the same behaviour 1 or 2 weeks later. Thus, the DBN, which was implemented from MAESTRO, can serve as a predictive tool in the sense that it estimates the probability of contamination phenomena occurring. Despite the innovative initiative, the resulting conditional probabilities did not show any relevant shellfish production area that could consistently explain the neighbouring area in the next week. Furthermore, it was shown that phytoplankton dispersion presents a higher positive relation from West to East. Patrício et al. (2022) [19] 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 [20, 21]. 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 [22, 23].

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) [24] 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.

1.3 Objectives

The growing interest in causality algorithms to explore life sciences [25–27] has led to the development of several state-of-the-art time series causality discovery algorithms, including Granger Causality (GC) [28], Temporal Causal Discovery Framework (TCDF) [29], Peter-Clark Momentary Conditional Independence (PCMCI) [30], Vector Autoregressive Linear Non-Gaussian Acyclic Model (VAR-LiNGAM) [31], and Dynamic Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (DYNOTEARS) [32]. In this thesis is desired to explore the potential of these causal discovery algorithms to identify the environmental variables with the most significant influence on biotoxins concentration in bivalve molluscs and evaluate which are more suitable.

Therefore, there are several stages in this process. The first objective was the evaluation of causality algorithms through benchmark datasets. As with most real problems, we don't know the ground-truth of this complex problem (contamination of bivalve molluscs), and since it's not possible to recreate a complex real system in its entirety, the causality algorithms were tested on several real datasets with known ground-truth. Thus, the algorithms were tested on real-world problems with different levels of complexity, to determine which method would be most suitable for our study.

The second objective regards the acquisition of the multiple time series collected from 2015 to 2020 in the Portuguese coast. The data was gathered from two different sources, namely IPMA and Coper-

nicus. IPMA provides weekly *in-situ* monitoring of 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 Portuguese coast is divided into 13 coastal zones and 28 estuarine-lagoon zones, making 41 shellfish production areas, as illustrated in Figure 1.1. Meteorological data is also compiled, including temperature, wind and rainfall. Copernicus [33] is the Earth observation component of the European Union's space programme, which provides diary *chl-a* concentration and SST measured by remote sensing (satellite data).

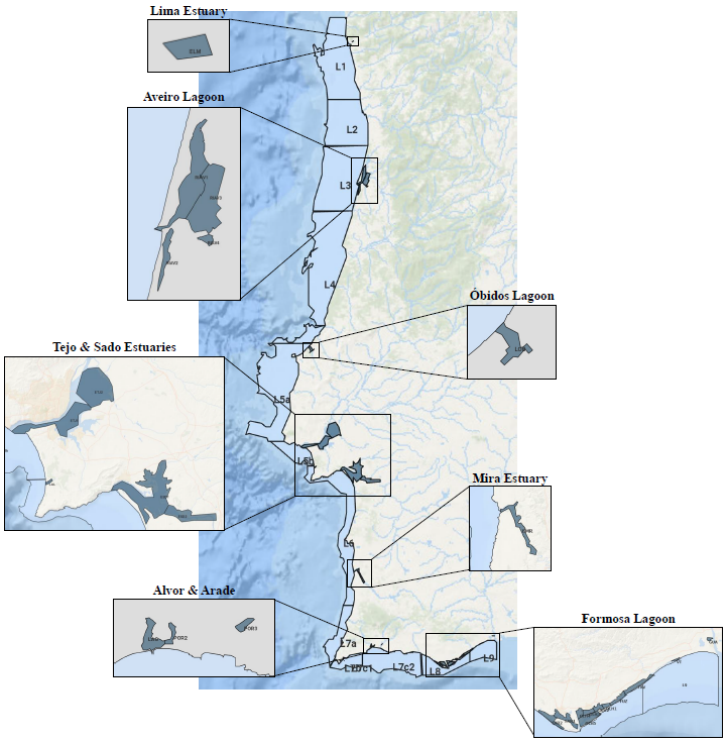


Figure 1.1: Shellfish production areas across the Portuguese coast. Adapted from [34]. The boundaries of production areas are defined by IPMA based on their geographical and coastal characteristics, as well as historical data on local catches in each maritime jurisdiction [12].

The third objective refers to preprocessing the acquired data, namely data integration (integration of multiple time series with the same sampling time), data cleaning (correction of typing errors), data selection (choosing only relevant production areas for the study), data imputation (semi-automatic imputation method to fill the missing data) and data transformation (stationarity and normalization).

Finally, the fourth objective was to choose and employ the causal discovery algorithm with the best performance on the benchmark datasets to obtain causal associations between DSP biotoxin contamination and the biological (DSP toxins-producing phytoplankton), oceanographic (SST and *chl-a*) and meteorological (maximum temperature, wind intensity, wind direction and rainfall) variables

This way, a novel perspective for inferring the relationships that govern biotoxin accumulation is proposed, which can aid in decision-making and improve seafood safety regarding bivalve mollusc consumption in Portugal.

1.4 Contributions

Based on the objectives set out in Section 1.3, this thesis makes several contributions. Firstly, it provided a review of theoretical concepts on causality and causality algorithms applied to time series and a review of past work associated with the MATISSE project. The second contribution was the development of a framework for evaluating various causality methods and, based on the results obtained, choosing the discovered causality algorithm to apply in our case study. Finally, based on the method chosen, we determined the causal relationships between variables that play an active role in shellfish contamination.

Some of the above findings resulted in the following approved article in an international 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 at *24th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*).

1.5 Thesis Outline

This thesis is organised as follows. Chapter 2 introduces the theoretical concepts of causality, correlation, graphical causal models, structural causal models and causality discovery algorithms applied to time series data. Chapter 3 presents and describes two benchmark datasets for temporal data, in order to validate the causal algorithms and determine the hyperparameters that perform best. Chapter 4 presents a real-world problem, the contamination of bivalve molluscs on the Portuguese coast. All the steps involved in data preprocessing, the selection of shellfish production areas and the results achieved by the method that performed best on the experimental data are carefully described. Finally, in Chapter 5, we denote the main conclusions and suggestions for future work.

Chapter 2

Background

This chapter is divided into four sections. The first introduces the concept of causality, correlation and the significant areas of research behind causality, namely discovery causality. The second section presents causal models, their main characteristics and how they can be graphically represented. The third section denotes the framework employed by the causality discovery algorithms. Finally, the fourth section introduces concepts of time series, the adaptation of causality to time series and the state-of-the-art methods for time series causality discovery.

2.1 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.

In recent years, the exponential growth of datasets and technological advances has given artificial intelligence an increasingly active role in various scientific fields. Typical machine learning algorithms have produced outstanding results in certain tasks, such as prediction and classification. However, due to the lack of transparency, some open questions remain about how decisions are made. Causality arose to solve some of the current machine learning limitations (correlation-based) and clarify how decisions are made (black boxes).

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.

2.2 Causation vs Correlation

Understanding the distinction between inferring correlation and inferring causation is fundamental. Theoretically, correlation is just a statistical measure of the association between variables in which two variables are correlated when they show a similar trend. Causation means a cause-and-effect relationship, which implies that changes in one variable cause changes in the other.

The difference between these concepts is so important that the sentence: "correlation is not causation" is taken as gospel truth. Although a causal link may exist when there is a substantial correlation between occurrences, two events occurring consecutively do not imply there is a causal relationship. In the literature, when two variables are highly correlated and seem to be causally connected, but the relationship is not causal, it is known as spurious correlation. Figure 2.1 represents one example of spurious correlation. According to the figure, the yearly number of films Nicolas Cage appears in correlates with the yearly number of people who drowned in a pool. Does Nicolas Cage have a clause in his contract that obliges him to accept more films for the sake of the number of drownings? Or does Nicolas Cage encourage people to watch his movies by a pool? A coincidence appears to be the better answer.

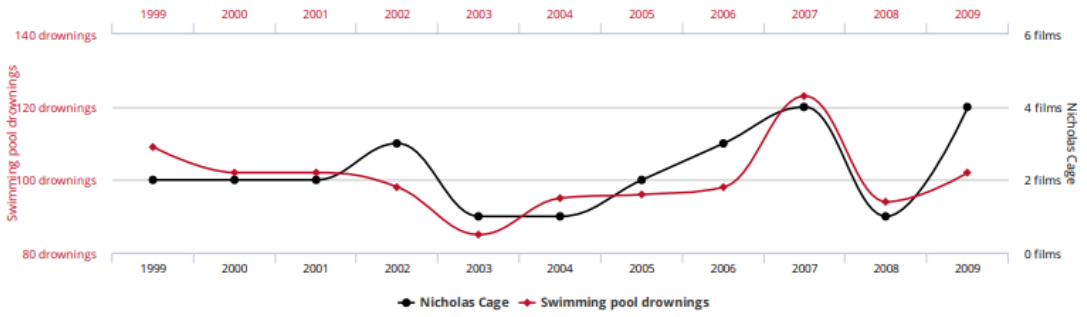


Figure 2.1: The yearly number of people who drowned by falling into a pool correlates with the yearly number of films Nicolas Cafe appeared in. Adapted from [35].

Spurious correlation might lead to untruthful and inaccurate conclusions, reducing the model's effectiveness. In addition to coincidental associations, there are two main reasons for spurious correlation: the third variable problem and the directionality problem [36]. The former happens when two effects are generated by the same cause (confounder). Even though they are two separate events, they share information about the initial event, which can lead to a false sense of causality. The latter occurs when two variables are correlated, but it is impossible to distinguish the cause from the effect.

2.3 Graphical Causal Models

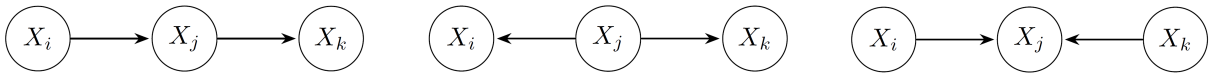
A causal graph $\mathcal{G} = (X, \mathcal{E})$ is a direct graph defined by a set of d nodes $X = \{X_1, \cdots, X_d\}$ and a set of edges $\mathcal{E}$, that represent a joint distribution [37]. Each node is associated with a unique variable and any two consecutive nodes connected by an edge $(i, j) \in \mathcal{E}, (i \neq j)$, are called adjacent nodes, where (i, j) represent an edge between X_i and X_j. To represent causal relationships between variables, only direct edges will be used, denoted by $X_i \to X_j$ to express that X_i has a causal effect on X_j. Furthermore, we say that X_i is a parent of X_j and X_j is a child of X_i. The set of parents of X_j and the set of children of X_i are written as $\mathcal{P}a_j^{\mathcal{G}}$ and $\mathcal{C}h_i^{\mathcal{G}}$, respectively.

A directed path in $\mathcal{G}$ is a sequence of nodes $X_1, \cdots, X_d$, in which for any two adjacent nodes, X_i and X_j, there is a directed edge $X_i \to X_j$ with consistent arrowhead direction such that $(i, j) \in \mathcal{E}$ but $(j, i) \notin \mathcal{E}$ [38, 39]. If a graph only contains a directed acyclic path, there is no directed path that starts and ends at the same node, and therefore $\mathcal{G}$ belongs to the well-known Directed Acyclic Graphs (DAGs). We will consider that the causal structure is based only on DAGs and therefore, satisfies the causal Markov condition, such that we can factorize the joint distribution, $P(X)$, as stated in the following recursive decomposition [40]:

$$P(X) = \prod_{i=1}^{d} P(X_i | \mathcal{P}a_i), \tag{2.1}$$

where $P(X_i|\mathcal{P}a_i)$ is the conditional distribution of X_i given its parents [41, 42]. This Equation is equivalent to the local Markov assumption, and in fact, is one of the fundamental grounds for the Bayesian networks, which allows us to understand the conditional independence embedded in the causal graph and how the variables are related. However, this assumption only gives us information about the independencies in a DAG. Therefore, the Minimality assumption is often used to make the local Markov assumption "stronger" [35]. In addition to the local Markov assumption, it is assumed that adjacent nodes are dependent in a DAG. For instance, if X_i causes X_j, from now on, we know that X_j must change in response to changes in X_i, meaning that they are associated (statistically dependent).

To formalize the (in)dependences from observational data into DAGs there are three basic building blocks: chain, fork and immorality, presented in Figure 2.2. Let us suppose two edges (i, j) and $(j, k) \in \mathcal{E}$ and three different nodes X_i, X_j and $X_k \in X$.



(a) Chain (b) Fork (c) Immorality or v-structure

Figure 2.2: Basic building blocks. Adapted from [37].

In the chain (Figure 2.2a), while $X_i \to X_j$ and $X_j \to X_k$ are dependent, X_i through its influence on X_j, causally affects X_k. In the fork (Figure 2.2b) and immorality (Figure 2.2c), there is no causation between X_i and X_j. In the former, X_i and X_k are at least partially associated because they have a confounder, i.e, a variable, X_j, that is a common cause of both variables. In the second, X_j is a collider node, i.e, a variable that is caused by two unrelated variables, which means that there is no causation

nor association between X_i and X_k. In practice it makes sense, it is not because two events generate the same effect that they have some association. A flood breaking a window due to high rainfall does not share any expected relationship with a missile breaking a window, yet both generate the same effect.

More types of edges are presented in the literature, such as undirected, bidirected and partially directed, which can give different interpretations to the graph. For instance, we say that a relationship is undirected when causation is detected, but it is impossible to infer who is the causal parent between two variables ($X_i \to X_j$ or $X_i \leftarrow X_j$). Bidirected edges refer to associated variables, and so it is assumed there is an unmeasured confounder. Finally, in partially directed relationships, either X_i is a causal parent of X_j, or there is an unmeasured confounder.

Some additional assumptions are often inferred to model observational data in causal discovery methods. Most methods in the field rely upon the causal sufficiency assumption, which states that for all variables in the DAG, there are no confounders unobserved [35, 40]. When this occurs, this set of variables is said to be causally sufficient. Although, sometimes this condition does not hold, and therefore some methods try to work around the existence of latent variables[1].

About the notion of d-separation, two nodes X_i and X_k are d-separated by a set of nodes Z if all the DAG's paths between X_i and X_k are blocked by Z, which can be denoted by $X_i \perp\!\!\!\perp_{\mathcal{G}} X_k | Z$. We say a path is blocked by a set of nodes Z if there exists a node X_j in the path such that one of the following conditions is satisfied [39, 43]:

1. $X_j \in Z$ and X_j is a non-collider node, that is:

   (a) $X_i \to X_j \to X_k$ or

   (b) $X_i \leftarrow X_j \leftarrow X_k$ or

   (c) $X_i \leftarrow X_j \to X_k$.

2. $X_j \notin Z$ nor any descendant and X_j is a collider node.

However, if there is at least one unblocked path between X_i and X_k, they are said to be d-connected. Furthermore, the d-separation property implies conditional independence, so according to the global Markov assumption, if X_i and X_k are d-separated in the DAG, $\mathcal{G}$, conditioned on $Z (X_i \perp\!\!\!\perp_{\mathcal{G}} X_k | Z)$, then they are independent in the distribution, $\mathcal{P}$, conditioned on $Z (X_i \perp\!\!\!\perp_{\mathcal{P}} X_k | Z)$:

$$X_i \perp\!\!\!\perp_{\mathcal{G}} X_k | Z \Rightarrow X_i \perp\!\!\!\perp_{\mathcal{P}} X_k | Z, \tag{2.2}$$

Moreover, the faithfulness assumption allows us to go the other way around, i.e., from independences in the distribution to infer d-separations in the DAG [35]

$$X_i \perp\!\!\!\perp_{\mathcal{G}} X_k | Z \Leftarrow X_i \perp\!\!\!\perp_{\mathcal{P}} X_k | Z, \tag{2.3}$$

[1] Variables inferred indirectly using observable variables

2.4 Structural 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 [37]:

$$X_i := f_i(\mathcal{P}a_i^{\mathcal{G}}, \mathcal{U}_i), \qquad i = \{1, \cdots, n\}, \tag{2.4}$$

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. Figure 2.3 illustrates an SCM consisting of a set of equations M and the corresponding causal graph.

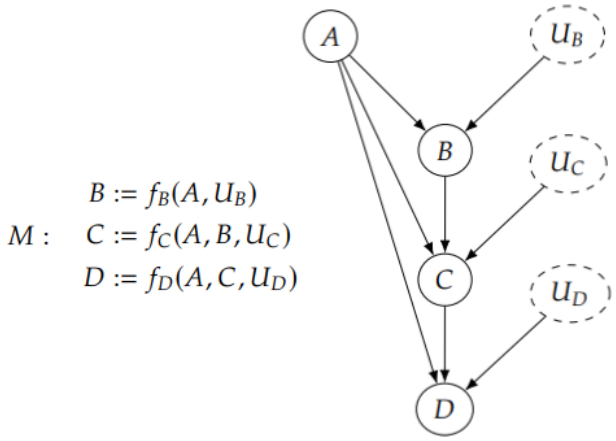


Figure 2.3: Causal graphs and respective structural equations. Adapted from [35].

2.5 Causal Discovery and Time Series Analysis

Generally speaking, there are two types of real-world data: cross-sectional and time series data [40]. Cross-sectional data are characterised by not considering temporal differences between observations, so data is collected without any particular sequential order. This presents a major disadvantage for causality analysis. Since they represent a single point in time, exploring temporal causal precedence is impossible. On the other hand, time series data are a sequence of observations recorded sequentially through time, and there may be a correlation between distinct observations. In this case, due to the time component, we can assume temporal causal precedence, i.e., events in the present cannot cause events in the past.

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. Figure 2.4 explores the potential of these methods, accounting for the most common challenges. The chosen algorithm must be able to adapt to the data and be able to interpret linear and/or non-linear relationships. Additionally, one must distinguish direct causes from indirect causes and spurious correlations. For instance, the presence of the common driver X^2 ($X^1 \leftarrow X^2 \rightarrow X^4$) induces a spurious link between X^1 and X^4, and the indirect path $X^3 \rightarrow X^2 \rightarrow X^1$ induces a spurious link between X^3 and X^1.

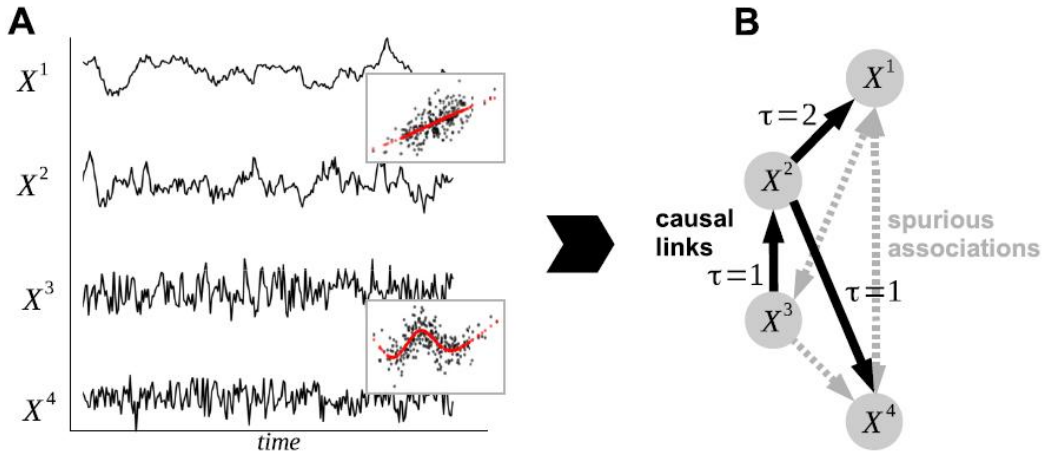


Figure 2.4: Causal network reconstruction (B) from a multivariate time series (A). The solid black arrows represent true causal links, while the dashed grey arrow represents spurious associations. Adapted from [44].

In Figure 2.5b, the so-called summary causal graph illustrates the temporal causal graph. There are at least three representations used in the literature: full time causal graph (2.5a), window causal graph (2.5b) and summary causal graph (2.5c). Consider a multivariate time series Y_t expressed by a causal graph $\mathcal{G} = (X, \mathcal{E})$. Due to the time component, the set of vertices X comprises the collection of components $y_t^1, \cdots, y_t^n$ at each time t. Two variables $X_{t-\tau}^i$ and X_t^j are directed connected through an edge of $\mathcal{E}$ in a full time causal graph if and only if X^i causes X^j at time t with time lag of $0 < \tau$ for $p = q$ and with a time lag of $0 \leq \tau$ for $p \neq q$. Note that any temporal representation assumes temporal priority (mentioned before as temporal causal precedence), which means that all causes must occur before its effect to establish a causal connection.

The window causal graph is adopted to reduce the window size, which is now given by the largest time gap between causes and effects. While the full time causal graph uses all time instants, which is unrealistic for extended time series, the window causal graph only covers a fixed number of timestamps. This reduction from full time causal graph to window causal graph is only possible via the consistency throughout time property, also known as causal stationarity [44, 45]. A causal graph $\mathcal{G}$ is said to be consistent throughout time if and only if all causal links $X_{t-\tau}^i$ and X_t^j remain constant in direction over

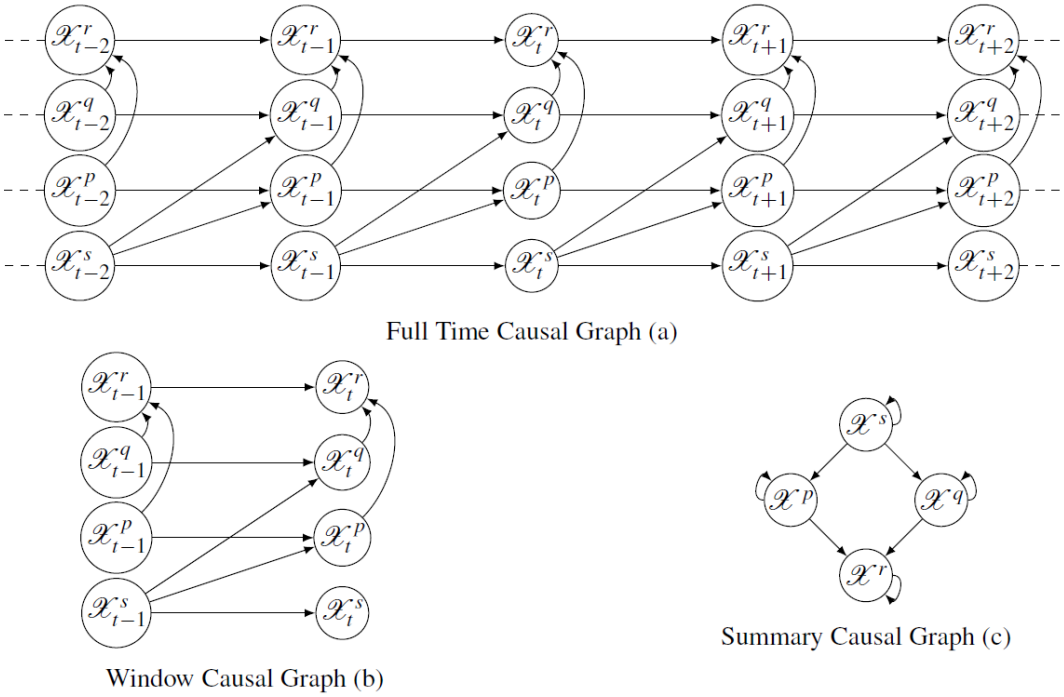


Figure 2.5: Causal graph representations: full time causal graph (a), window causal graph (b) and summary causal graph (c). Adapted from [45].

time in the graph:

$$X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t | X^-_t \ \{X^j_t\} \text{ holds for all } t, \tag{2.5}$$

Finally, we can compress the causal graph even further. In most real-world problems, it is difficult to determine which time lag is the cause of another, which makes it hard to detect causal relationships between time instants. Thus, the summary causal graph emerged to represent causal relations among time series without any time information.

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.

2.5.1 Granger Causality

One of the oldest and most well-known concepts to infer causal relationships from time series data is Granger causality (GC) [28]. This statistical concept verifies whether a time series can be used to improve the prediction about another time series, respecting temporal precedence [46].

Pairwise Granger Causality Consider two univariate time series y^p and y^q, we say that y^p Granger causes y^q if y^p contains unique and statistically significant information on the past observations of y^q that is not available in y^q, which can be formalized using a vector autoregressive (VAR) topology [45]:

$$y^q_t = a^q_0 + \sum_{\tau=1}^{\tau_{\max}} a^q_\tau y^q_{t-\tau} + \epsilon^q_t, \tag{2.6}$$

$$y_t^q = a_0^q + \sum_{\tau=1}^{\tau_{\max}} a_\tau^q y_{t-\tau}^q + \sum_{\tau=1}^{\tau_{\max}} a_\tau^p y_{t-\tau}^p + \epsilon_t^q, \tag{2.7}$$

where $\tau_{\max}$ represents the maximum time lags τ, a_τ^q and a_τ^p are the contributions of lagged observations of y_q and y_p, respectively, and ϵ_t denotes an independent noise with zero mean. The Equation 2.6 is named *restricted model* because it only uses past observations of y^q to predict its current values. The *full model* (Equation 2.7) trough the past values of both time series, y^q and y^p predict the present values of y_q. In this case, a_τ^p will be non-zero, and therefore we can say y^p Granger causes y^q with lag τ, denoted by $y_{t-\tau}^p \rightarrow y_t^q$. To determine whether the *full model* is "statistically significant" relative to the *restricted model*, statistical tests are conducted. The Fisher F-test (F) is a common choice to handle a hypothesis test, considering that y^p does not Granger cause y^q as the null hypothesis [47]:

$$F = \frac{\mathrm{RSS}_1 - \mathrm{RSS}_2/p}{\mathrm{RSS}_2/(n-2p-1)}, \tag{2.8}$$

where n is the time series length, and RSS_1 and RSS_2 are the full and restricted model's residual sum of squares, respectively. Figure 2.6 shows an example of a time series y^p that Granger causes y^q. It is possible to identify that temporal precedence is fulfilled, since the effects always occur after the cause, with time lag τ . Furthermore, the Granger concept is verified, considering that y^p has unique information that improves the prediction of y^q. Although pairwise analysis does not capture Granger's original idea that all appropriate information should be incorporated into the analysis, it can also be applied in a multivariate context. However, numerous spurious correlations will likely result without the ability to condition common dependencies (confounders).

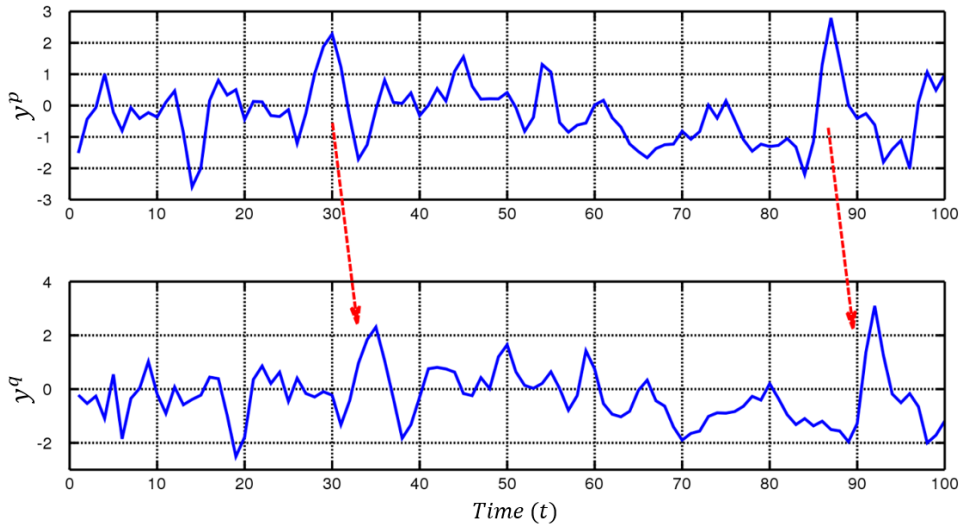


Figure 2.6: Example of a time series y^p Granger causing a time series y^q. Adapted from [48].

Multivariate Granger Causality The Multivariate Granger Causality (MVGC) is an extension of the previous method, to solve the spurious correlation problems. The MVGC intends to condition common dependencies, based on the original Granger idea. In this way, the *full model* in a multivariate settings (Equation 2.10) uses all time series while the *restricted model* (Equation 2.9) does not use y^p in

the VAR model:

$$y_t^q = a_0^q + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{\tau=1}^{\tau_{\max}} a_\tau^r y_{t-\tau}^p + \epsilon_t^q, \tag{2.9}$$

$$y_t^q = a_0^q + \sum_{r=1}^{d} \sum_{\tau=1}^{\tau_{\max}} a_\tau^r y_{t-\tau}^r + \epsilon_t^q, \tag{2.10}$$

The main disadvantage pointed to Granger causality is the lack of ability to deal with real-world causal relationships. However, determining causal relationships is valuable for interpreting and improving prediction models. Given the interpretability of the model, many approaches have been developed to deal with these problems. In some cases, Granger causality is used as a feature selection tool in the sense that it selects the most relevant features from the dataset to be later worked on by a given algorithm [47, 49].

2.5.2 Temporal Causal Discovery Framework

Deep learning models have been studied to overcome the limitations of Granger causality [50]. The explainability given by the concept of Granger allied with the emergence of deep neural networks allows us to investigate further nonlinear causal analysis and explore high computational demand [51–53]. Here, we explore a deep-learning extension of MVGC: the Temporal Causal Discovery Framework (TCDF) [54] 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 (Figure 2.7). In addition, TCDF models autocorrelation by using the target time series as input.

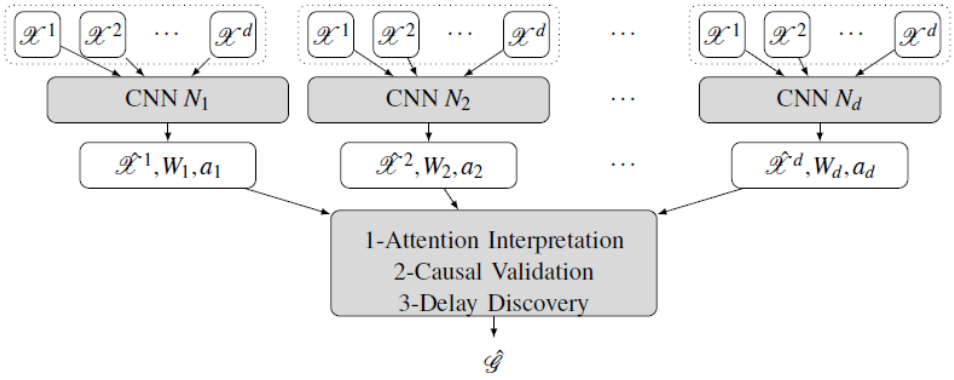


Figure 2.7: TCDF neural network. Adapted from [45].

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.

In short, a temporal causal graph (step4) is presented after (Step 1) Time series prediction, (Step 2) Attention Interpretation, (Step 3(a)) Causal Validation and (Step 3(b)) Delay Discovery, which is illustrated in Figure 2.8.

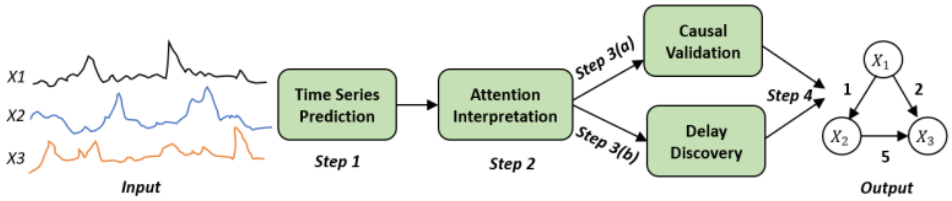


Figure 2.8: TCDF neural network. Adapted from [55].

The biggest disadvantage of TCDF is the difficulty in tuning the hyperparameters: kernel size, dilation coefficient, hidden layers, number of epochs, learning rate and loss function. Another difficulty is due to the maximum time delay incorporated by the model. The number of timestamps read by the sliding kernel increases if the number of hidden layers, dilation coefficient, or kernel size increases.

2.5.3 Peter-Clark Momentary Conditional Independence

The Peter-Clark Momentary Conditional Independence (PCMCI) algorithm introduced by Runge et al. (2019) [30], is a temporal extension of the Peter-Clark (PC) algorithm [56]. This constraint-based approach aims to exploit the conditional independencies embedded in the data in order to build the underlying causal graph. To achieve this, several assumptions must be followed: causal Markov condition, causal sufficiency, causal stationarity and faithfulness.

Peter-Clark Algorithm The Peter-Clark (PC) algorithm [56], named after its authors Peter Spirtes and Clark Glymour, is a state-of-the-art constraint-based method for causal discovery. Based on statistical procedures for determining conditional independences, the PC algorithm seeks to learn causal relationships from non-temporal datasets. Considering the graph in Figure 2.9a as the true graph, i.e., the graph that represents some underlying structure (also known as ground truth), the procedure followed by the PC algorithm is illustrated in Figure 2.9, and the steps are described below:

1. Creation of a complete undirected graph. This DAG skeleton represents the edges between every pair of variables, which means we are not assuming any conditional independencies about the distribution, resulting in Figure 2.9b.

2. Identify the skeleton: Conditional independence tests are performed to identify the skeleton. Considering the size of the conditioning set, edges between independent variables are removed [35].

(a) Start with the empty conditioning set $X \perp\!\!\!\perp Y\{\}$ and test pairwise independencies. For instance, $A \perp\!\!\!\perp B$. The only path between A and B is blocked by the collider C, and so, there is no association flowing between A and B. In Figure 2.9c, the edge $A - B$ is removed.

(b) Independence tests with conditioning set of size one. All pairs of variables are conditioned on their adjacent variables. In Figure 2.9d, all edges that are not connecting with C are removed because all pairs of variables are conditionally independent given C.

(c) Continue to increase the conditioning set size until there are no further statistical decisions to make. In the considered example, the previous step already removed every single edge that was conditional independent and discovered the skeleton of the true graph.

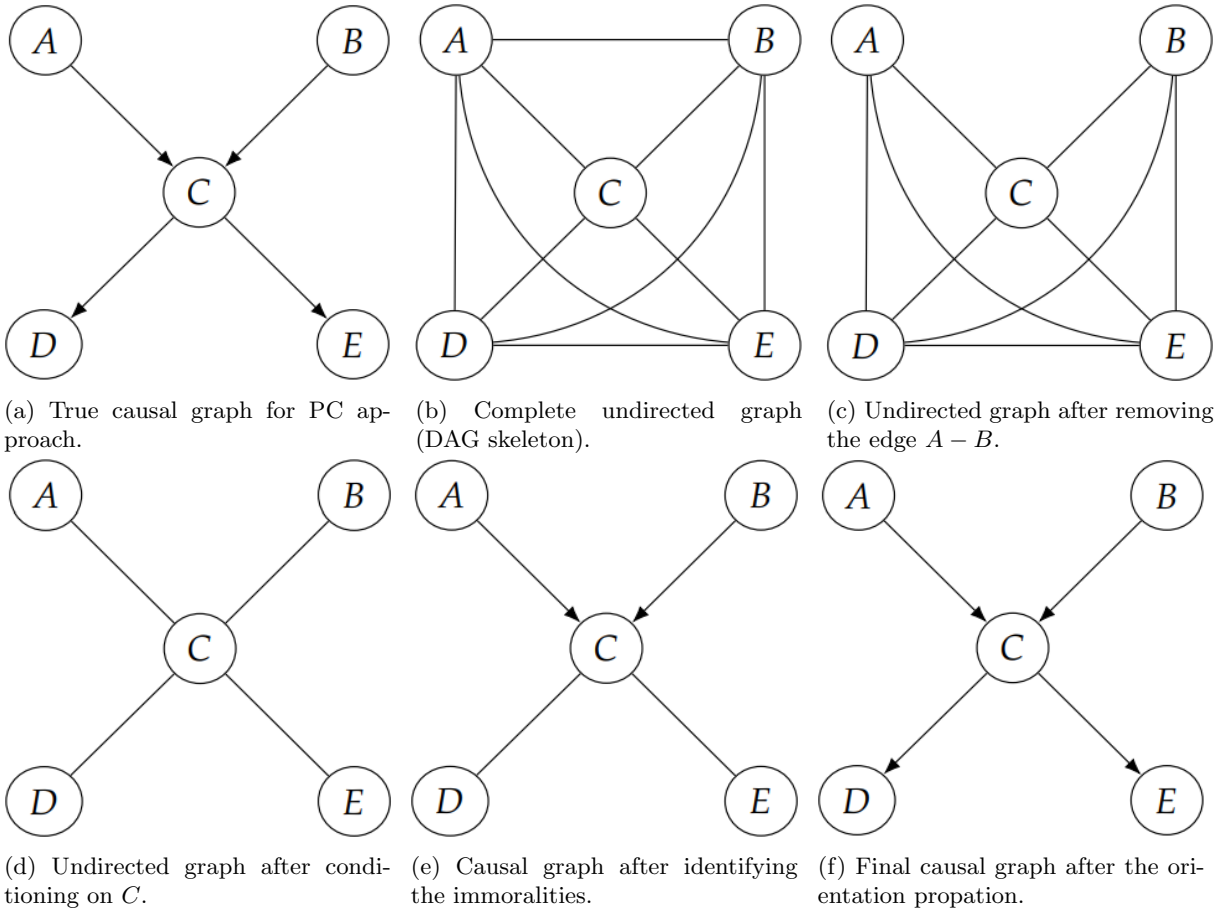
3. Identifying the immoralities: For any triple $X - Z - Y$, if Z is not on the conditioning set that makes X and Y conditionally independent, and there is no edge between X and Y, the only way to causally represent the triple is through an immorality structure. Therefore, Figure 2.9e presents the triple $A - B - C$ by an immorality: $A \rightarrow C \leftarrow B$.

4. Orienting qualifying edges incident on colliders: Assuming that all the immoralities were discovered in the former step, for each triple such that $X \rightarrow Z - Y$, where there is no edge between X and Y, any edge $Z - Y$ can be oriented as $Z \rightarrow Y$ (known as orientation propagation). Figure 2.9f orient the two remaining edges after the orientation propagation. Note that guiding $D \rightarrow C$ (or $E \rightarrow C$) would not be possible, as an immorality would be generated. This cannot happen since it is considered that all immoralities are determined in the previous step.

It should also be noted that although all causal relationships were determined in this exercise, sometimes the causal graph is not fully oriented because it is not possible to distinguish chains from fork structures. Thus, some edges may remain undirected in the final causal graph, which means that although two variables are known for sharing association, it is not possible to infer the direction the edge points. In this situation, the PC algorithm, instead of returning only the final DAG, will return all Markov Equivalence graphs [57]. Different graphs belong to the same Markov Equivalence Class if they have the same statistical interpretation, i.e., encode the same conditional independencies. Figure 2.10 illustrates an example of a Markov Equivalence Class where the conditional independence $A \perp\!\!\!\perp B|\{C\}$, $A \perp\!\!\!\perp D|\{C\}$ and $B \perp\!\!\!\perp D|\{C\}$ signifies the same for all.

In short, the PC algorithm is a powerful tool for determining causal links in non-temporal data through independent conditional tests. However, it presents some limitations, namely high computational demand (due to the size of the conditioning set, number of variables and sample size), lack of ability to deal with possible unobserved/latent confounders (faithfulness assumption assumes that all causes are present in the dataset) and depends on the conditional independence test being compatible with the distribution of the data, otherwise it may not detect some edges or even introduce spurious links.

PC Temporal Extension with MCI Tests The PCMCI algorithm was designed to solve some of the weaknesses associated with the PC algorithm, mainly the lack of a temporal interpretation. Therefore, this time series causal discovery algorithm was architect to be able to detect time-lagged linear and



(a) True causal graph for PC approach.

(b) Complete undirected graph (DAG skeleton).

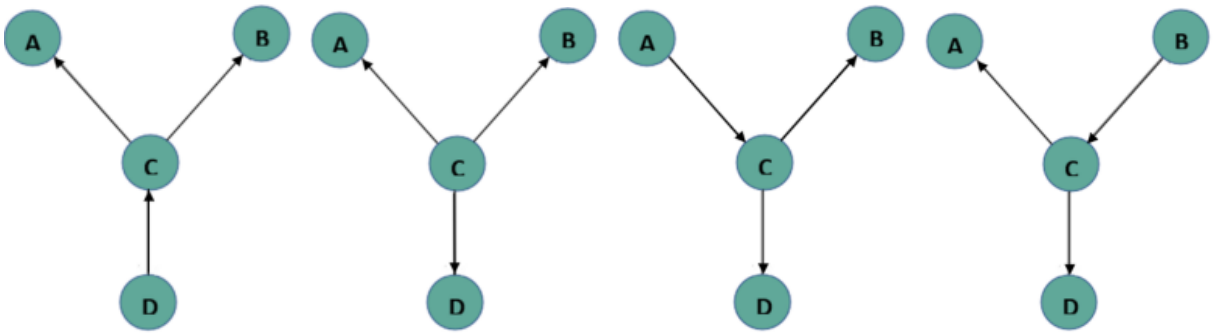
(c) Undirected graph after removing the edge $A - B$.

(d) Undirected graph after conditioning on C.

(e) Causal graph after identifying the immoralities.

(f) Final causal graph after the orientation propation.

Figure 2.9: Example of the PC algorithm steps. Adapted from [35].



(a) True causal graph.

(b) Example 1.

(c) Example 2.

(d) Example 3.

Figure 2.10: Markov Equivalence Class representation. Adapted from [57].

nonlinear causal relations in a window causal graph. It can be framed into two main phases: the PC_1 phase [30, 44] and the Momentary Conditional Independence (MCI) phase [58, 59].

In the first stage, PC_1 is a variant from the previously presented PC algorithm [56] in its more robust modification called PC-stable [60]. Runge et al. (2018, 2019) [30, 44] took a general formulation for cross-sectional variables and adapted it to time series. For connections between time series, temporal order can be used to provide an orientation rule naturally. The new approach is presented below. Similarly to the PC algorithm in which we start with a complete undirected graph, here we start by considering that

Table 2.1: Conditional independence test and corresponding assumptions used in the study case analysis.

Conditional independence test	Assumptions
Partial Correlation (ParCorr) test	Univariate, continuos variables with linear dependencies and Gaussian noise
Gaussian Processes and Distance Correlation (GPDC) test	Univariate, continuos variables with additve dependencies

all variables $X_t^j \in X_t$ belong to the preliminary set of parents up to a lag $\tau_{\max}$ [61]:

$$\hat{\mathcal{P}}(X_t^j) = (X_{t-1}, X_{t-2}, \cdots, X_{t-\tau_{\max}}) \tag{2.11}$$

Considering the empty conditioning set ($p = 0$), if the null hypothesis $X_{t-\tau}^i \perp\!\!\!\perp X_t^j$ can not be rejected at a significance threshold α_{pc}, a variable $X_{t-\tau}^i$ is removed from $\hat{\mathcal{P}}(X_t^j)$ and p is increased to $p = 1$. For the remaining iterations ($p > 0$), if the null hypothesis (Equation 2.12) cannot be rejected, the edge is removed from $\hat{\mathcal{P}}(X_t^j)$ and the p is increased iteratively to $p = p + 1$ for any subsequent iteration:

$$X_{t-\tau}^i \perp\!\!\!\perp X_t^j | S \quad \text{for any } S \text{ with } |S| = p, \tag{2.12}$$

where S are the subset of the strongest p parents which iterates through different combinations of subsets $\hat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}$. In each iteration, after the conditional independence test (Table 2.1), the set of parents $\hat{\mathcal{P}}(X_t^j)$ are sorted by the absolute test statistic value. The algorithm converges for a causal link $X_{t-\tau}^i \to X_t^j$ whenever the null hypothesis described in Equation 2.13 is rejected at a significance threshold α_{pc} and $S \subseteq \hat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}$.

$$X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \hat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}, \tag{2.13}$$

Note that PC_1 stands out from the PC algorithm by conditioning on the p parents with stronger dependencies instead of testing all combinations of conditions. Additionally, the chosen conditional independence test, such as the chosen significance level, can affect the assessment of the strength of each causal edge. Table 2.1 presents the conditional independence tests that were used in the case study analyzed later. While the partial correlation (ParCorr) test aims to determine linear cause-effect relationships, the Gaussian processes and distance correlation (GPDC) test allows you to determine non-linear causal relationships.

Finally, in the second phase, a Momentary Conditional Independence test (MCI) is performed. This stage is motivated by the information-theoretic measure momentary information transfer [58, 59]. Considering the previously selected parents $\hat{\mathcal{P}}(X_t^j)$ as conditions, all variables pairs $(X_{t-\tau}^i, X_t^j)$ are tested and a causal link $X_{t-\tau}^i \to X_t^j$ is established, if and only if [30]:

$$X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j | \hat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}, \hat{\mathcal{P}_{px}}(X_{t-\tau}^i), \tag{2.14}$$

where the time delay $\tau \in \{1, \cdots, \tau_{\max}\}$, $i, j \in \{1, \cdots, N\}$ and $\hat{\mathcal{P}_{p_x}}(X_{t-\tau}^i) \subseteq \hat{\mathcal{P}}(X_t^j)$ denotes the p_x

strongest parents. Since the conditioning set is based on both parents of $X^i_{t-\tau}$ and X^j_t, one account for autocorrelation and make the MCI test statistic a measure of causal strength.

Regarding instantaneous relations, PCMCI can not support them and, therefore, consider undirected contemporaneous links for $\tau = 0$. More recently, Runge et al. (2020) [62] introduced the PCMCI+ to integrate instantaneous relations by leading separately the edge removal for lagged conditioning sets and instantaneous conditioning sets.

2.5.4 Vector Autoregressive Linear Non-Gaussian Acyclic Model

The VAR-LiNGAM introduced by Hyvarianen et al. (2010) [31] seeks to identify the information contained in the noise of the variables and based on this information, identify the direction and strength of the causal relationships. In particular, this noise-based approach is a temporal extension of LINGAM [63].

LiNGAM (Linear Non-Gaussian Acyclic Model) Considering a linear, acyclic structural equation model, with non-Gaussian error terms, the causal model in the two-variable case can be described as [45]:

$$X^p = \xi^p, \tag{2.15}$$

$$X^q = a^{p,q} X^p + \xi^q, \tag{2.16}$$

where ξ^p and ξ^q are non-Gaussian noise and $X^p \perp\!\!\!\perp \xi^q$. Under this distribution, X^p has a causal effect on X^q, which means that X^q receive the noise information provided by ξ^p, through X^p. On the other hand, X^p does not learn any information on ξ^q, resulting in an asymmetry that LiNGAM seeks to explore.

Note that without the aforementioned assumptions, it would not be possible to identify the direction of the causal relationship and distinguish cause from effect. Let's consider two linear variables X and Y, that is, $Y = X + \varepsilon$, where $X \perp\!\!\!\perp \varepsilon$. Figure 2.11 illustrates the scatter plots of the variables X and Y in columns 1 and 3, and in columns 2 and 4, the predictor and residuals from two regression tasks.

Three configurations of X and ε are expressed in the three rows: (case 1) Both Gaussian, (case 2) Uniformly distributed and (case 3) Super-Gaussian distribution. In the last two scenarios, X and ε are non-Gaussian, resulting in an asymmetry between X and Y, only when regression of X given Y (anti-causal direction). Thus, we can conclude that the predictors are independent of the residuals only for the correct causal direction. To uncover the underlying asymmetries, the model response is reframed as follows:

$$X = AX + \xi, \tag{2.17}$$

where X is a vector of observed variables, A is a strictly lower triangular matrix and ξ is a vector of noise. The matrix A that represents the causal directions between the variables of X can be determined

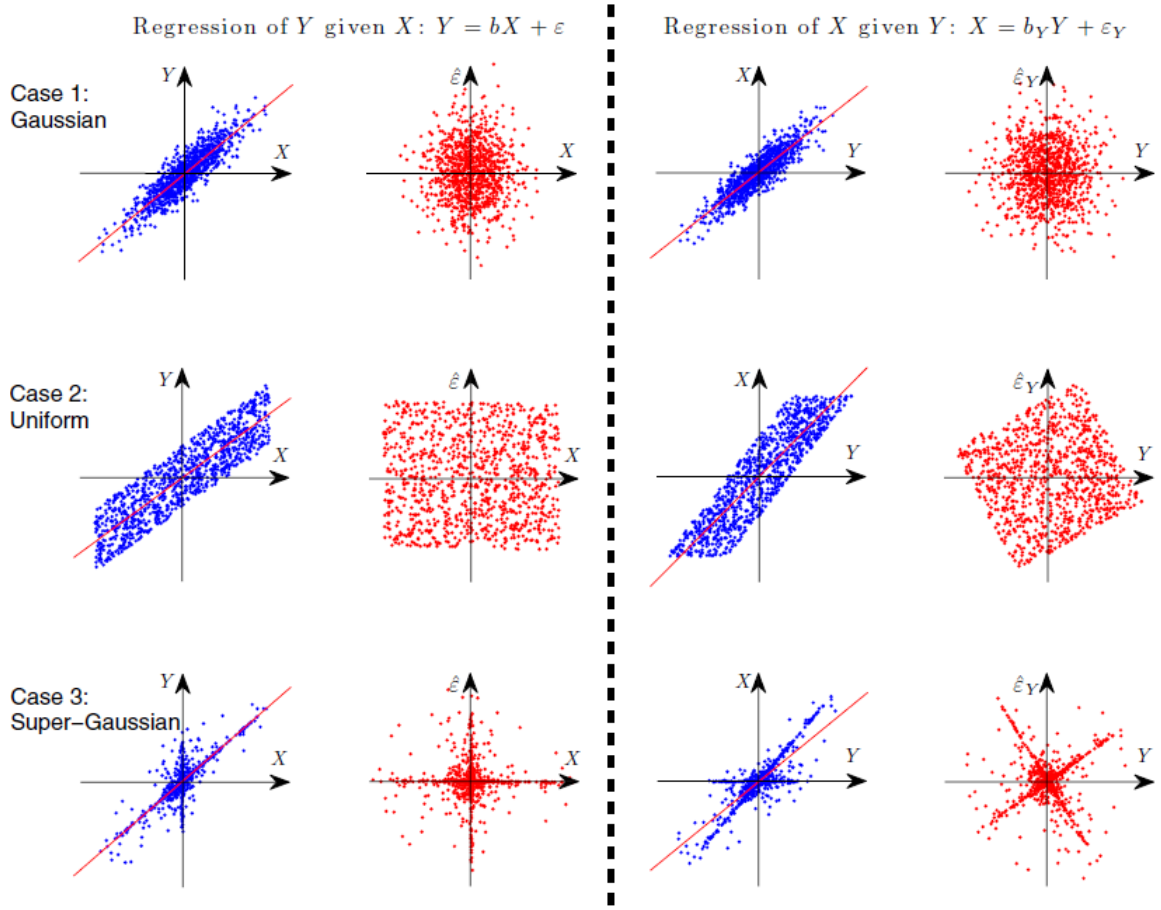


Figure 2.11: Causal asymmetry example between two variables with linear relations. Adapted from [38].

by rearranging the previous equation into:

$$X = B\xi, \tag{2.18}$$

$$B = (I - A)^{-1}, \tag{2.19}$$

Initially, the mixing matrix B was determined through an independent component analysis (ICA) and then, the matrix A were computed. More recently, Shimizu et al. (2011) [64] employed the DirectLiNGAM, an improved extension of the original algorithm. This method constructs an auto-regressive model and recursively determines the causal hierarchy based on the independence between the residuals and predictors [65]. For that, each variable acts as a predictor and the residuals are obtained by applying that predictor to other variables. The most independent predictor from the residuals of its target variables is ranked highest in the causal order. To remove the effects of the earlier determined predictors, the same analysis is executed on those residuals in each subsequent step. The sequence obtained is used to construct a strictly lower triangular A establishing the direction of causation, and a covariance-based regression, such as least squares, estimates the causal effects, i.e., the strength of the connections $A_{i,j}$. Finally, an Adaptive Lasso [66] is utilised to prune A and get a sparser causal model, by using the l_1 penalty.

LiNGAM Temporal Extension By considering the variables over a time window defined by the maximum time lag $\tau_{\max}$, LiNGAM can be extended to VAR-LiNGAM, based on a structural vector autoregressive that can be formulated as:

$$X_t = \sum_{\tau=0}^{\tau_{\max}} A_\tau X_{t-\tau} + \xi_t, \tag{2.20}$$

where the matrix A_τ for $0 < \tau \leq \tau_{\max}$ correspond to lagged effects and A_0 correspond to instantaneous effects. Furthermore, the VAR-LiNGAM is founded on three premises: (i) ξ_t are non-Gaussian, (ii) Over time, ξ_t are among themselves mutually independent and temporally uncorrelated, and (iii) The matrix A_0 is an acyclic graph [67].

The following approach can implement the VAR-LiNGAM [31]. First, the model is rewritten as a vector autoregressive model to the causal relationships among variables with a time delay $\tau > 0$:

$$X_t = \sum_{\tau=1}^{\tau_{\max}} M_\tau X_{t-\tau} + \xi_t, \tag{2.21}$$

The residuals of the prediction of X_t are obtained through a least square practice. Secondly, a LiNGAM analysis on these residuals is conducted, to compute the instantaneous causal model A_0. For $\tau > 0$, the A_τ parameters are achieved by parametrization of M_τ:

$$A_\tau = (I - A_0)M_\tau, \tag{2.22}$$

2.5.5 Dynamic Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning

The DYNOTEARS [32] is a dynamic extension of the Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS) [68], by incorporating temporal dynamics. This score-based approach seeks to replace conditional independence tests with a defined scoring criterion. Based on Dynamic Bayesian networks (DBN), this causal discovery task pretends to find a causal graph that maximises a scoring metric of how well the data fits a given graph. Additionally, it 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. Figure 2.12 illustrates the inter-slice (dash-lines) and intra-slice (solid-lines) from a DBN composed of $d = 3$ nodes and autoregression order $p = 2$. Only causal links with time instant t are represented by a black arrow.

Consider a time series with M independent realizations and $\{x_t^m, t = 0, \cdots, T\}$ be the m^{th} time series component at time t, for $x_t^m \in \mathbb{R}^d$ (where d represents the number of variables/nodes). This can be formulated by using a SVAR model:

$$x_t^{mT} = x_t^{mT} W + x_t^{mT} A_1 + \cdots + x_{t-P}^{m\ T} A_p + z_t^{mT}, \tag{2.23}$$

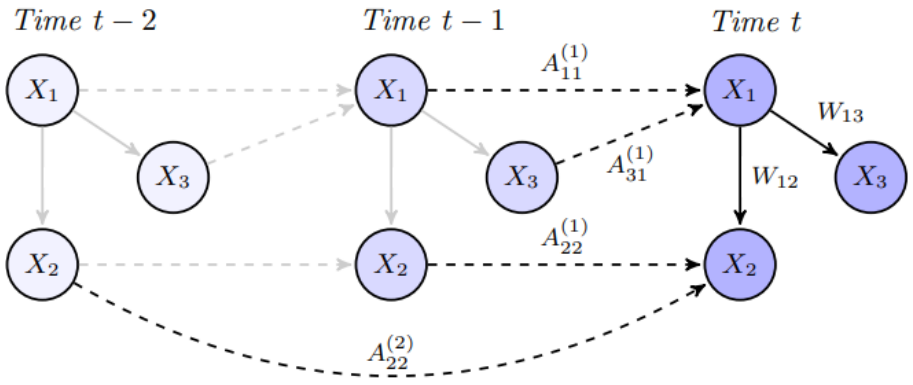


Figure 2.12: Explanation of DBN inter-slice (dash lines) and intra-slice (solid lines) dependencies. Adapted from [32].

where p is the autoregressive order, $t \in \{p, \cdots, T\}$, $m \in \{1, \cdots, M\}$, $z_t^{m\,T}$ is a vector of noise terms and, W and $A_1, \cdots, A_p$ correspond to a weighted adjacency matrice of intra-slice and inter-slice edges, respectively.

Considering that the network is constant over time, W is acyclic, and the matrices parameterize a DBN, the latter equation can be formulated in a matrix form as:

$$X = XW + Y_1 A_1 + \cdots + Y_p A_p + Z, \tag{2.24}$$

where $Y_1, \cdots, Y_p$ are time-lagged version of X, X is an $n \times d$ matrix whose rows are $x_t^{m\,T}$. Finally, let $A = [A_1^T | \cdots | A_p^T]^T$ and $Y = [Y_1 | \cdots | Y_p]$ be the $pd \times d$ matrix of inter-slices weights and $n \times pd$ matrix of time-lagged data, respectively.

$$X = XW + YA + \cdots + YA + Z, \tag{2.25}$$

This general formulation permits to include, in the lagged data matrix Y, only the time instant that has a causal impact at time t. That is not required to cover the continuous sequence of time slices from $t - p$ to $t - 1$.

The search in the DAG search space for the optimal weighted adjacency matrices W and A must take into account that W must be acyclic for the network to be acyclic as well. A does not create cycles, since the edges only go forward in time. This optimization problem can be formulated by minimizing the least squares loss with the acyclic constraint [69]:

$$\begin{aligned} \min_{W,A} \quad & \frac{1}{2n} \|X - XW - YA\|_F^2, \\ \text{s.t} \quad & W \text{ is acyclic}, \end{aligned} \tag{2.26}$$

where $\|.\|_F^2$ is the squared Frobeniusnorm and n is the sample size (or number of n rows) described as $n = M(T + 1 - p)$. Lastly, l_1 penalties can be further integrated into the objective function to enforce

sparsity in W and A:

$$\begin{aligned} \min_{W,A} \quad & \frac{1}{2n}\|X - XW - YA\|_F^2 + \lambda_W\|W\|_1 + \lambda_A\|A\|_1, \\ \text{s.t} \quad & W \text{ is acyclic,} \end{aligned} \tag{2.27}$$

where $\|.\|_1$ represents the element-wise l_1 norm. From a computational point of view, DYNOTEARS presents four hyperparameters to be tuned: the regularization constants $lambda_W$ and $lambda_A$, and the weight thresholds τ_W and tau_A. The weight thresholds aim to eliminate weights close to 0 by thresholding the entries of W and A.

2.5.6 Software Implementation and Related Work

Table 2.2 summarizes some of the information on the previous causal discovery algorithms, namely the papers that gave rise to the methods, the Python libraries that can be used to replicate the algorithms and some of the articles in which they are applied.

Table 2.2: Python implementations.

Method	Library	Referenced Articles
Pairwise GC [28]	statsmodels: https://github.com/statsmodels/statsmodels	[70–75]
TCDF [54]	tdcf: https://github.com/M-Nauta/TCDF	[45, 55, 76]
PCMCI/PCMCI+ [30, 62]	tigramite: https://github.com/jakobrunge/tigramite	[25, 44, 77–82]
VAR-LiNGAM [83]	lingam: https://github.com/cdt15/lingam	[45, 84, 85]
DYNOTEARS [32, 39]	causalnex: https://github.com/quantumblacklabs/causalnex	[65, 86, 87]

The use of these algorithms on real-world data has been increasing, primarily due to the improvement in the decision-making power they can introduce. Real-world data is not easily modulated, which makes the task much more difficult. Faced with challenges such as non-linear distributions, increased sample sizes and different sampling rates, many authors introduce innovations to these state-of-the-art methods.

Amornbunchhornvej et al. (2019) [70] proposed a variable-lag Granger causality to infer causal relationships in time series with arbitrary lags. Typical approaches strongly assume that all effects occur with a fixed time delay, which is not always accurate in real-world problems. In this sense, the authors propose to align time series using Dynamic Time Warping (DTW) and subsequently apply Granger causality to determine causal relationships with arbitrary dynamic time delays.

Papagiannpoulou et al. (2017) [75] developed a nonlinear Granger causality algorithm based on random forests to discover the relationships between climate and vegetation. Due to its excellent computational scalability, this method replaced the traditional autoregressive and linear Granger causality models. Thus, the model using random forests was applied to multidecadal satellite records to determine the predictors of monthly vegetation variability on a global scale, which could predict 14% more variability of vegetation anomalies.

Ohmura (2022) [84] explores the potential of the VAR-LiNGAM in detecting the causal relationship between stock price trends and ruling party support in Japan. The results showed that stock price indirectly causally affects the ruling party approval rate through the consumer confidex index. In addition, the consumer confidex index was confirmed to be a lagged effect of the shared prices and the ruling

approval rate an instantaneous effect of the consumer confidex index.

Einizade et al. (2022) [85] propose CGP-LiNGAM, a version of VAR-LiNGAM that aims to reduce the number of parameters and model VAR coefficient matrices with graph polynomial filters. Thus, the authors concluded that CGP-LINGAM has significantly fewer parameters than VAR-LiNGAM and ensure that with just one underlying causal graph, it allows the detection of instantaneous and time-lagged causal relationships and has a more robust performance against a low number of time samples and a high degree of causal dependence.

Howard et al. (2023) [65] explores ten temporal causal discovery algorithms in autonomous driving benchmarks, including pairwise granger causality, TCDF, PCMCI, VAR-LINGAM and DYNOTEARS. It is noteworthy that DYNOTEARS was the method with the best F1-score, while TCDF and VAR-LiNGAM were the methods with the worst performance.

Menegozzo1 et al. (2020) [81] explores the potential of PCMCI to reveal causal relationships in simulated realizations that reflect the characteristics of an ultra-processed food manufacturing industry process. The PCMCI with the CMIKnn (Nearest-neighbor permutation test based on conditional mutual information) independence test was able to detect non-linear relationships. In particular, 50% of the causal links were well identified without any FPR.

Alvarez et al. (2022) [82] applies PCMCI and PCMCI+ to soil moisture (SMOS) and vegetation dynamics (NDVI) of tropical South America and the effects of the Eastern Pacific (EP) and Central Pacific (CP) El Niño-Southern Oscillation (ENSO) events. The existence of bidirectional nonlinear links between SMOS and NVDI was determined. Additionally, NVDI showed a weaker nonlinear temporal persistence than soil mixture. Furthermore, ParCorr detected more causal links than PCMCI: $7 > 6$ with CP as target variable $and 5 > 3$ for EP.

Kretschmer et al. (2016) [78] investigated possible Arctic mechanisms which to understand northern hemisphere mid-latitude extreme winters in Eurasia and North America. Based on the power of detection of PCMCI, was confirmed that Arctic sea ice extent in autumn is an important driver of winter circulation in the mid-latitudes.

Furthermore, several authors thoroughly describe these temporal causality discovery algorithms, some focusing more on the theoretical component and others using benchmark datasets for comparison with these state-of-the-art algorithms [39, 40, 45, 50].

Chapter 3

Experimental Analysis

This chapter develops several data simulations employing the causal discovery algorithms introduced in Section 2.5 and can be divided into four sections. The first section describes the different sources and datasets utilised, presenting some examples of ground truths. The second section introduces the evaluation metrics for the different causal methods. The third section explains the parameter settings for the various methods and experiments. Finally, the fourth section discusses the best algorithm, given the different simulations in time series benchmark datasets. All the algorithms employed were executed using *Python*.

3.1 Materials Description

As mentioned, causal discovery seeks to identify and quantify causal relationships from observational data. However, we can not validate the data without prior knowledge about the system's behaviour. In supervised problems, such as prediction, the next predicted value is compared with the real one. In contrast, in unsupervised problems, such as causality, one either has prior knowledge about the interaction between the variables or accepts the results. As no causal graph represents the phenomenon of bivalve contamination, benchmark data will be used to validate the models. The outstanding advantage of benchmark data is that we have the corresponding causal temporal graph (ground truth) for a given time series. Thus, we can directly compare the graph generated by the causal model and the ground truth.

Benchmark data can be synthetic or real. Synthetic benchmark data allows the shape of the characteristics of the data as desired by varying the sample size, the number of variables, linearity, the time lag between causal relationships, and how autocorrelated each time series is, among others. However, even modelling several datasets will not be possible to fully reproduce the interaction between the variables of the real mechanism. The presence of hidden confounders, autocorrelation, Gaussian noise or not and the type of structures used (pair, fork, v-structure) are some of the associated uncertainties.

The real benchmark data, on the other hand, is closer to reproducing the characteristics of our dataset because it already has the mentioned uncertainties "built-in". These data correspond to causal graphs validated by scientists and researchers and, therefore, are real data with known ground truth. In this

case, we can not control the linearity of the data or the time interval between causal relationships. Still, we can choose datasets with a sample size and number of variables similar to our system's. This way, two real benchmarks of several time series data were chosen.

Blood-oxygen-level dependent functional magnetic resonance imaging (BOLD FMRI) Blood-oxygen-level dependent functional magnetic resonance imaging (BOLD FMRI) is the first real-world benchmark dataset introduced [88]. This benchmark 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. The original data can be accessed through https://www.fmrib.ox.ac.uk/datasets/netsim/index.html, and a preprocessed version is available at https://github.com/M-Nauta/TCDF/tree/master/data/fMRI.

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.) [89]. 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. In addition, some time series were pre-processed so that the sample size stayed within 300. Bivariate analysis may not be the most complex from the point of view of the number of variables, but it does present very specific challenges, such as linearity and noise. It was assumed that no selection bias, confounding, or feedback loops existed. The original data can be accessed through https://webdav.tuebingen.mpg.de/cause-effect/.

Table 3.1 provides the characterisation of the time series from the BOLD FMRI and CauseEffect-Pair.Taking into account Section 2.5, the ground truth is represented as a windowed causal graph and a summarised causal graph, for BOLD FMRI and CauseEffectPair benchmarks, respectively.

3.2 Setup

With the presence of benchmark data, it is essential to determine which causality algorithms best adapt to the underlying mechanisms of the data and which are the best hyperparameters. The choice of hyper-parameters can seriously affect the model's performance, so it is advisable to test various values to get the best of the algorithms. Table 3.2 shows the hyperparameter values tested for each model, considering a $\tau_{\max} = 5$ for all methods.

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

Table 3.1: Selected time series from the BOLD FMRI and CauseEffectPairs benchmarks.

BOLD FMRI			CauseEffectPairs			
Sim.	# Nodes	Sample size	Pair	Variable1	Variable 2	Sample size
1	5	200	0001	Altitude	Temperature	349
2	10	200	0002	Altitude	Precipitation	349
3	15	200	0003	Longitude	Temperature	349
8	5	200	0004	Longitude	Sunshine hours	349
10	5	200	0020	Longitude	Temperature	349
11	10	200	0021	Longitude	Precipitation	349
12	10	200	0048	Indoor Temp.	Outdoor Temp.	168
13	5	200	0049	Ozone Conc.	Temperature	365
14	5	200	0050	Ozone Conc.	Temperature	365
15	5	200	0051	Ozone Conc.	Temperature	365
16	5	200	0053	Ozone Conc.	Wind, Radiation, Temp.	365
17	10	200	0077	Temperature	Solar Radiation	365
18	5	200	0081	Temperature	CO2 flux (BE-Bra)	365
21	5	200	0082	Temperature	CO2 flux (DE-Har)	365
22	5	200	0083	Temperature	CO2 flux (US-PFa)	365
23	5	200	0093	Precipitation	Runoff	432
24	5	200	-	-	-	-

$\in$ [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.

Table 3.2: Causal discovery algorithms and the corresponding hyperparameters. The experimental name is to facilitate the analytical description of the results obtained.

Method	Experimental Name	Hyperparameters
GC	GC	Test static: F-Test, $\alpha = 5\%$
TCDF	TCDF	$K = c = \tau_{\max} + 1$, Epochs=[1000,2000,3000,4000], $\lambda = [0.001, 0.005, 0.01, 0.05, 0.1, 0.5]$, Optimizer=[Adam, RMSprop], s =[0, 0.2,0.4,0.6,0.8,1]
PCMCI	1	ParCorr, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5]
	1C	ParCorr, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5], corrected p-values
	3	GPDC, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5]
	3C	GPDC, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5], corrected p-values
PCMCI+	1P	ParCorr, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5]
	1PC	ParCorr, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5] , corrected p-values
	3P	GPDC, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5]
	3PC	GPDC, α_{PC}=[0.01,0.05,0.1,0.2,0.3,0.4,0.5], corrected p-values
VAR-LiNGAM	VL	Criterion=None, Prune=True, α_{limit}=[0,0.05,0.1,0.15,0.2]
DYNOTEARS	DYN	$\lambda_w \in [0.01, 0.1]$, $\lambda_a \in [0.1, 1, 10]$, $w_\tau \in [0.01, 0.1]$

3.3 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. Figure A.2 represents a confusion matrix between the actual class and the predicted class, 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.

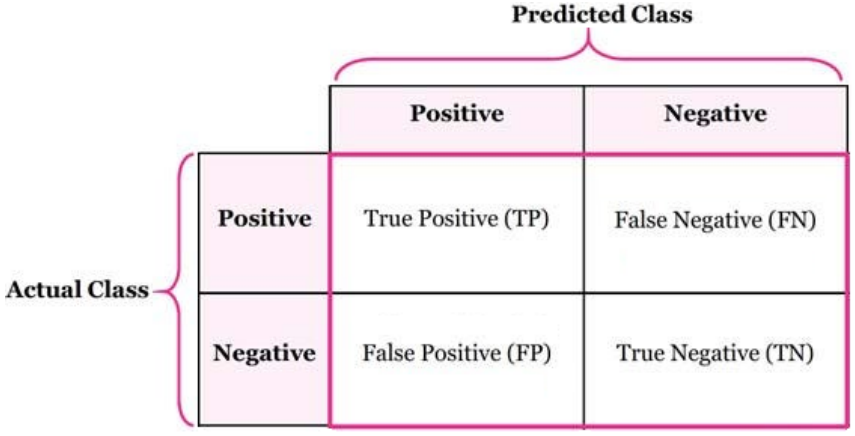


Figure 3.1: Confusion matrix representation. The matrix entries cover the number of causal links (oriented edges) predicted by the causal algorithm to exist or not in the ground truth.

Considering the previous notation (TP, FP, TN, FP) for binary classification, the precision, recall, F1-score (F1) and false positive rate (FPR) are obtained:

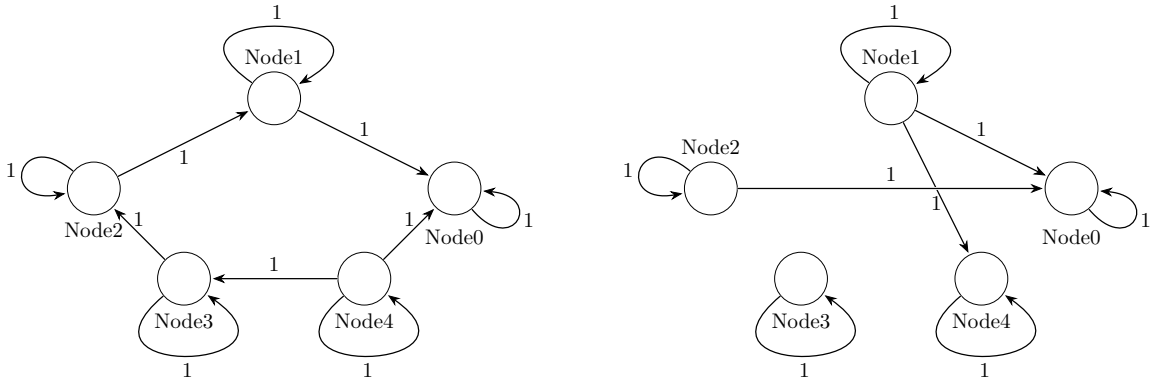
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.2}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{3.3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TP} + \text{FP}}, \tag{3.4}$$

As an example, Figure 3.2 illustrates the ground truth of simulation 22 of the BOLD FMRI benchmark and the corresponding learned graph by the 3P algorithm with $\alpha_{\text{PC}} = 0.01$ (3.2b). The manipulated algorithm presents: 6 TP (Node1 → Node0, Node1 → Node1, Node0 → Node0, Node2 → Node2, Node3 → Node3, Node4 → Node4), 2 FP (Node1 → Node4, Node2 → Node0), 4 FN (Node3 → Node2, Node2 → Node1, Node4 → Node3, Node4 → Node0, 13 TN: Node2 → Node3, Node3 → Node4, Node0 → Node4, Node0 → Node1, Node1 → Node2, Node1 → Node3, Node3 → Node1,Node3 → Node0, Node0 → Node3, Node4 → Node2, Node2 →Node4, Node4 → Node1, Node0 → Node2).



(a) Ground truth.

(b) Learned graph: 3C algorithm with $\alpha_{\text{PC}} = 0.01$.

Figure 3.2: Comparison between ground truth (a) and learned graph (b) from simulation 22 from BOLD FMRI benchmark.

In addition, Table 3.3 shows an F1-score of 0.67 for the previously represented model and calculates the metrics for the various hyperparameters tested: precision, recall, F1-score and FPR. This example was also chosen to clarify how the best hyperparameter is selected. If several hyperparameters obtained the highest F1-score, such as $\alpha_{\text{PC}} = 0.3$, $\alpha_{\text{PC}} = 0.4$ and $\alpha_{\text{PC}} = 0.5$, both would be considered and counted as the best hyperparameter in this simulation unless the FPR of one of them was lower than that of the others. For example, if 0.67 was the highest F1-score, between $\alpha_{\text{PC}} = 0.01$, $\alpha_{\text{PC}} = 0.05$, and $\alpha_{\text{PC}} = 0.1$, the hyperparameter $\alpha_{\text{PC}} = 0.1$ would be chosen because it had the lowest FPR. Additionally, if all the hyperparameters resulted in an F1-score with the same value in a given experiment, no combination of hyperparameters would be counted because no model fitted the data better than the others. In the end,

the hyperparameters of each experiment are added together, and the one with the highest count is chosen for a given model.

Table 3.3: Hyperparameter selection of the 3P algorithm for simulation 22 from BOLD FMRI benchmark.

	$\alpha_{\text{PC}} = 0.01$	$\alpha_{\text{PC}} = 0.05$	$\alpha_{\text{PC}} = 0.1$	$\alpha_{\text{PC}} = 0.2$	$\alpha_{\text{PC}} = 0.3$	$\alpha_{\text{PC}} = 0.4$	$\alpha_{\text{PC}} = 0.5$
TP	6	6	5	5	6	6	6
FP	2	2	0	0	1	1	1
FN	4	4	5	5	4	4	4
TN	13	13	15	15	14	14	14
Precision	0.75	0.75	1	1	0.86	0.86	0.86
Recall	0.60	0.60	0.50	0.50	0.60	0.60	0.60
F1	0.67	0.67	0.67	0.67	0.71	0.71	0.71
FPR	0.13	0.13	0.0	0.0	0.07	0.07	0.07

3.4 Experimental Results

This section demonstrates the steps followed to decide which method is most suitable for our case study. One of the big challenges when choosing a particular method is which hyperparameters to select. To this end, a grid search was conducted to determine which hyperparameters best represented the various datasets. Considering Table 3.2 and the different combinations that can be made with the hyperparameters, Table 3.4 represents the combination that offers the best F1-score.

The three best F1 scores are 0.88, 0.81, and 0.77, given by TCDF, 3P and 1P, respectively. The GC obtained the worst F1-score at 0.28, which is not surprising given that it is a bivariate version and was adapted to the multivariate context. Interestingly, a multivariate version of Granger causality in deep learning achieved the best F1-score. These results seem to be very promising. However, only a combination of hyperparameters is evaluated in real data without prior knowledge of the causal mechanism. It is impossible to perform a grid search to determine the best hyperparameters. Therefore, experiments were carried out with datasets with different particularities and degrees of complexity to determine which hyperparameters were used most frequently in real data to value their ability to adjust to the data presented. Figures 3.3a and 3.3b and Tables 3.5 and 3.6 show the number of times that a given hyperparameter was considered the best for a given experience. This analysis corresponds to the 1P, 3P, VAR-LiNGAM, DYNOTEARS and TCDF methods, respectively. GC was not explored as it presented much lower results than other methods and, therefore, would not be chosen for our case study.

In this way, 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 3.7 presents the results obtained using the previously mentioned parameters. Note that there was a significant reduction in the F1-score. 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 3.4: Combination of hyperparameters that provides the highest F1-score. Bold represents the highest F1-scores.

	GC	1	1P	1C	1PC	3	3P	3C	3PC	DYN	VL	TCDF
Sim.: 1	0.00	0.71	0.71	0.67	0.67	0.71	0.74	0.75	0.67	0.50	0.90	0.91
Sim.: 2	0.07	0.67	0.68	0.67	0.60	0.58	0.70	0.63	0.60	0.32	0.59	0.76
Sim.: 3	0.12	0.65	0.66	0.64	0.57	0.61	0.62	0.63	0.43	0.23	0.60	0.73
Sim.: 8	0.40	0.76	0.74	0.84	0.57	0.80	0.84	0.75	0.63	0.53	0.80	0.80
Sim.: 10	0.14	0.63	0.84	0.67	0.67	0.63	0.82	0.67	0.67	0.56	0.90	0.91
Sim.: 11	0.12	0.68	0.68	0.67	0.60	0.60	0.67	0.63	0.65	0.32	0.53	0.61
Sim.: 12	0.14	0.60	0.67	0.69	0.60	0.60	0.68	0.65	0.60	0.37	0.67	0.72
Sim.: 13	0.13	0.60	0.76	0.56	0.56	0.63	0.75	0.63	0.56	0.40	0.57	0.73
Sim.: 14	0.00	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.53	0.80	0.76
Sim.: 15	0.29	0.73	0.71	0.71	0.57	0.70	0.84	0.63	0.57	0.50	0.57	0.78
Sim.: 16	0.25	0.70	0.67	0.63	0.59	0.76	0.64	0.70	0.40	0.64	0.75	0.80
Sim.: 17	0.21	0.61	0.67	0.61	0.65	0.62	0.67	0.67	0.60	0.44	0.74	0.83
Sim.: 18	0.00	0.75	0.84	0.67	0.67	0.67	0.89	0.63	0.67	0.67	0.53	0.87
Sim.: 21	0.00	0.71	0.71	0.67	0.67	0.71	0.75	0.75	0.67	0.50	0.90	0.91
Sim.: 22	0.18	0.75	0.74	0.75	0.67	0.74	0.74	0.71	0.67	0.50	0.67	0.76
Sim.: 23	0.19	0.73	0.74	0.73	0.46	0.78	0.70	0.78	0.46	0.47	0.45	0.72
Sim.: 24	0.31	0.58	0.59	0.58	0.00	0.61	0.63	0.57	0.00	0.35	0.56	0.67
pair: 0001	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
pair: 0002	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
pair: 0003	0.00	0.00	0.67	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00
pair: 0004	0.00	0.00	0.67	0.00	0.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00
pair: 0020	0.00	0.00	1.00	0.00	0.00	1.00	0.67	0.00	0.00	1.00	0.00	1.00
pair: 0021	0.00	0.00	1.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	1.00
pair: 0048	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
pair: 0049	0.67	0.67	1.00	0.67	1.00	0.67	1.00	0.67	0.00	1.00	1.00	1.00
pair: 0050	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.67	1.00
pair: 0051	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.67	1.00
pair: 0053	0.40	0.43	0.43	0.33	0.25	0.40	0.40	0.40	0.40	0.57	0.50	0.67
pair: 0077	0.67	1.00	1.00	1.00	0.67	1.00	1.00	1.00	0.00	1.00	1.00	1.00
pair: 0081	0.67	0.67	0.67	1.00	0.00	0.67	0.67	0.67	0.00	1.00	1.00	1.00
pair: 0082	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
pair: 0083	0.67	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00	0.67	1.00
pair: 0093	0.00	1.00	1.00	0.67	0.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00
Final Mean	0.28	0.68	**0.77**	0.61	0.44	0.69	**0.81**	0.58	0.45	0.65	0.64	**0.88**

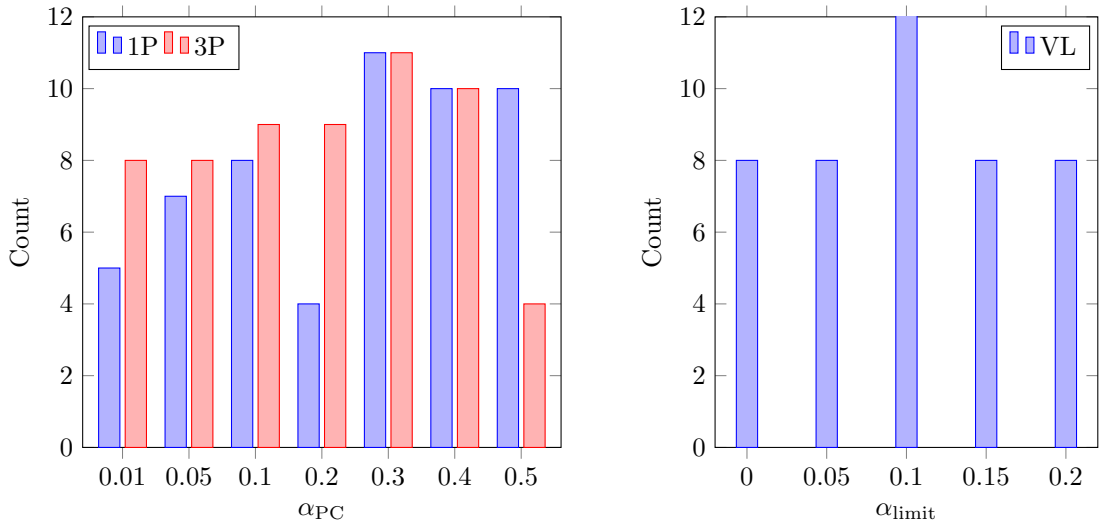
Table 3.5: Hyperparameter selection for the DYNOTEARS algorithm.

DYNOTEARS										
Hyperparameters	w_τ					λ_w		λ_a		
Values	0.1	0.2	0.3	0.4	0.5	0.01	0.1	0.1	1	10
Count	44	41	45	34	34	125	63	56	71	70

Table 3.6: Hyperparameter selection for the TCDF algorithm.

TCDF										
Hyperparameters	Epochs				λ					
Values	1000	2000	3000	4000	0.001	0.005	0.01	0.05	0.1	0.5
Count	60	67	65	59	33	32	40	40	49	57

Hyperparameters	s						Optimizer	
Values	0	0.2	0.4	0.6	0.8	1	Adam	RMSprop
Count	4	4	3	5	13	222	127	124



(a) 1P e 3P.

(b) VAR-LiNGAM.

Figure 3.3: Hyperparameter selection for the 1P, 3P and VL algorithms.

Table 3.7: 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

Chapter 4

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.

4.1 Data Preprocessing

This section presents a framework for collecting, processing and analyzing multivariate time series corresponding to each production area. Figure 4.1 illustrates the five stages of this process. Step 1 corresponds to data acquisition and integration, step 2 seeks to clean the data, step 3 explores and selects the pertinent data to be analyzed, step 4 fills the missing data with some criteria, and finally, step 5 manipulates the data to improve the final result. This order may change depending on the problem. In this case, it was represented as follows to facilitate the interpretation of the results.

Figure 4.1: Data pre-processing steps.

4.1.1 Data Acquisition and Data Integration

As mentioned, 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 series into a single 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 [9, 16, 90–93].

4.1.1.1 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. Figure 4.2 illustrates the position of the meteorological stations provided by IPMA along the Portuguese coast. For each production area, the nearest meteorological station is assigned using the station coordinates and the coordinates of the sampling points.

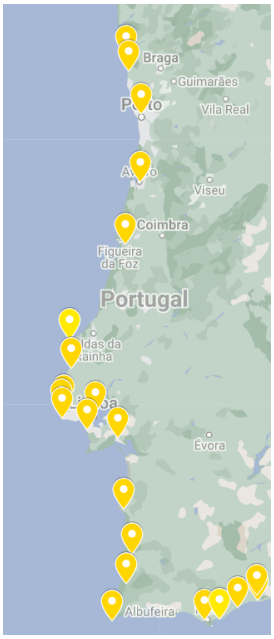


Figure 4.2: Meteorological stations provided by IPMA along the Portuguese coastline.

4.1.1.2 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 [33]. 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.

4.1.1.3 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. Table 4.1 describes the variables provided.

Table 4.1: Variables provided for each production area.

Variable	Description	Unit	Type
DSP toxins	DSP toxins concentration	μg AO equiv. kg^{-1}	Continuous
ASP toxins	ASP toxins concentration	mg DA equiv. kg^{-1}	Continuous
PSP toxins	PSP toxins concentration	μg STX equiv. kg^{-1}	Continuous
DSP phyto	DSP toxins-producing phytoplankton cell counts	cell L^{-1}	Continuous
ASP phyto	ASP toxins-producing phytoplankton cell counts	cell L^{-1}	Continuous
PSP phyto	PSP toxins-producing phytoplankton cell counts	cell L^{-1}	Continuous
SST	Mean sea surface temperature	K	Continuous
CHL	Mean of *chlorophyll-a* concentration	mg m^{-3}	Continuous
Min temp	Minimal air temperature	°C	Continuous
Mean temp	Mean air temperature	°C	Continuous
Max temp	Maximum air temperature	°C	Continuous
Wind int	Mean wind speed	m s^{-1}	Continuous
Card wind dir	Cardinal wind direction (N,S,E,$\cdots$)	-	Discrete
Wind dir	Mean wind direction	°	Continuos
Rainfall	Accumulated precipitation	mm	Continuous

4.1.2 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. Figure 4.3 shows some production areas (ELM, EMI, EMN1, EMR, ESD1, EDS2, ETJ1 and ETJ2) and the corresponding sampling points. As can be seen, many of the sampling points have writing 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

```
ELM
['Est. Lima - Jusante Ponte EiFel' 'Est. Lima - Jusante Ponte EiFe'
 'Est. Lima - Jusante Ponte Eiffel' 'Montante da Ponte Eiffel'
 'Jusante da Ponte' 'Jusante Ponte Eiffel' ' Jusante da Ponte Eiffel']
------------------------------
EMI
['Est. Minho _Vila Nova Cerveira' 'Est. Minho - Vila Nova Cerveira'
 'Est. Minho -Vila Nova Cerveira']
------------------------------
EMN1
['Murraceira Norte' 'Morraceira Norte']
------------------------------
EMN2
['Murraceira sul' 'Murraceira Sul' 'Entrada Braço Sul' 'Morraceira Sul'
 'Morraceira sul']
------------------------------
EMR
['Jusante da Ponte' 'Est. Mira' 'Est. Mira - Roncão' 'Est.Mira -Roncão'
 'Est. Mira – Casa Branca' 'a Roncão- Casa branca'
 'Est. Mira - Casa Branca' 'Roncão/Casa Branca' 'Estuário do Mira'
 'Estuario do Mira - Roncão' 'Roncão' 'Roncão - Casa Branca' 'Troviscais'
 'Frente ao ISN']
------------------------------
ESD1
['Est. Sado - Faralhão' 'Est. Sado - Canal da Vaia' 'Canal de Vaia'
 'Est. Sado - Gâmbia' 'stuário do Sado - Canal da Vai'
 'Est. Sado - Est. Marateca' 'Est. Sado' 'Canal da Vaia' 'Faralhão'
 'Mitrena' 'Zambujal' 'Palma']
------------------------------
ESD2
['Carrasqueira' 'Est. Sado - Palma' 'Est. Sado - Canal de Alcacer'
 'Est. Sado – Palma' 'Palma' 'Zambujal']
------------------------------
ETJ1
['Porto Brandão' 'Est. Tejo - Samouco' 'Est. Tejo - Baliza Ferro'
 'Samouco' 'Baliza de Ferro' 'Baliza Ferro' 'Base Aérea' 'Alcochete'
 'Cacilhas' 'Ponta dos Corvos']
------------------------------
ETJ2
['Baliza Ferro' 'Baliza de Ferro']
------------------------------
```

(a) Misspelled input errors.

```
ELM
['Est. Lima - Jusante Ponte Eiffel' 'Est. Lima - Montante Ponte Eiffel']
------------------------------
EMI
['Est. Minho - Vila Nova Cerveira']
------------------------------
EMN1
['Murraceira Norte']
------------------------------
EMN2
['Murraceira Sul' 'Entrada Braço Sul']
------------------------------
EMR
['Est. Mira - Jusante da Ponte' 'Est. Mira' 'Est. Mira - Roncão'
 'Est. Mira - Casa Branca' 'Est. Mira - Roncão/Casa Branca' 'Troviscais'
 'Frente ao ISN']
------------------------------
ESD1
['Est. Sado - Faralhão' 'Est. Sado - Canal da Vaia' 'Est. Sado - Gâmbia'
 'Est. Sado - Est. Marateca' 'Mitrena' 'Zambujal' 'Est. Sado - Palma']
------------------------------
ESD2
['Carrasqueira' 'Est. Sado - Palma' 'Est. Sado - Canal de Alcacer'
 'Zambujal']
------------------------------
ETJ1
['Porto Brandão' 'Est. Tejo - Samouco' 'Est. Tejo - Baliza Ferro'
 'Base Aérea' 'Alcochete' 'Cacilhas' 'Ponta dos Corvos']
------------------------------
ETJ2
['Est. Tejo - Baliza Ferro']
------------------------------
```

(b) Corrected misspelled erros.

Figure 4.3: Misspelled (a) and corrected (b) manual input errors.

Table 4.2: Sampling location and the most commercialized species in production areas with the most cases of DSP contamination. Note that the RIAV2 production area has different sampling points for the two species and, therefore, was divided into the codes RIAV2(M) and RIAV2(B), to facilitate the representation of mussel and cockle harvesting, respectively.

Code Name	Production Area	Sampling Location	Specie	Commercial Name
L1	Viana coast	Carreço	Mussel	Mytilus spp.
L2	Matosinhos coast	Leça da Palmeira	Mussel	Mytilus spp.
L3	Aveiro coast	Molhe da Vagueira	Mussel	Mytilus spp.
L5b	Cabo Raso-Lagoa de Albufeira coast	Caparica	Mussel	Mytilus spp.
L8	Faro-Olhão coast	Culatra	Donax Clam	Donax trunculus
L9	Tavira coast	Monte Gordo	Donax Clam	Donax trunculus
RIAV1	Aveiro ria	Triângulo das Correntes	Mussel	Mytilus spp.
RIAV2(M)	Aveiro ria	Docapesca Navio Santo Andre	Mussel	Mytilus spp.
RIAV2(B)	Aveiro ria	Costa Nova Sul da Ponte da Barra	Cockle	Cerastoderma edule

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.

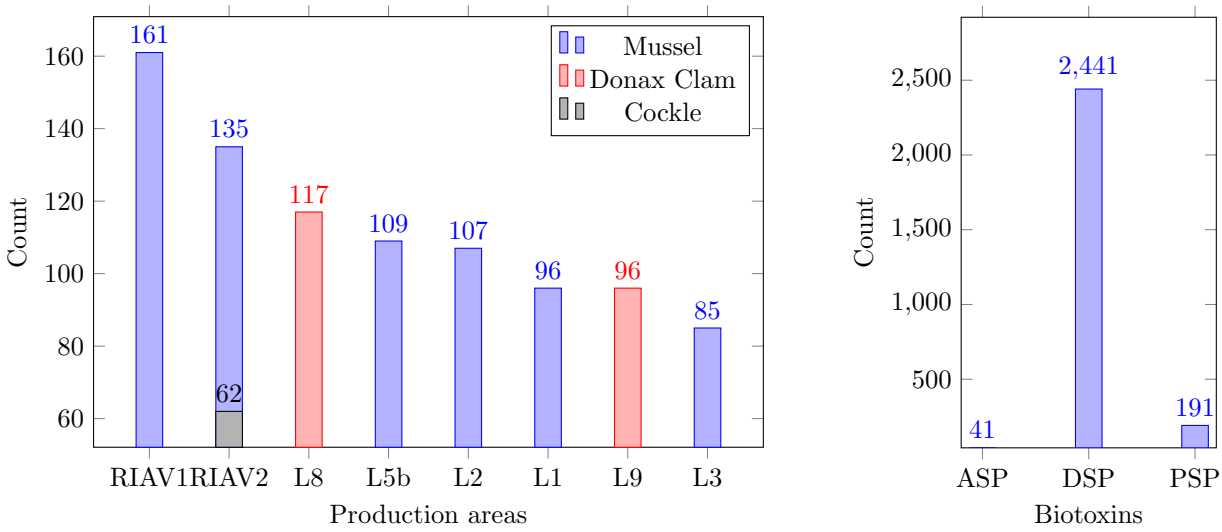
For the meteorological time series, IPMA assigned the code -990 to represent observations without measurements. Thus, the observations were replaced by missing data for later imputation.

4.1.3 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 with 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.

Figures 4.4b shows the number of registered cases where the ASP, DSP and PSP toxins exceeded the safety limits from 2015 to 2020. Given the values exhibited, only the DSP toxins makes sense to analyse due to a low number of toxic events. Note that the figures submitted represent all species and production areas. 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 4.4a represents the cases of DSP contamination with the most events per specie and Table 4.2 provides the sampling location and the most commercialized species for each production area. The estuarine-lagoon regions (RIAV1 and RIAV2) present higher toxicity levels than the coastal regions (remaining production areas).

Moreover, only three species were illustrated: mussels, donax clams and cockles. The fact that mussels



(a) Highest DSP concentration events above the safety limit per production area.

(b) Biotoxins concentration events above the safety limit.

Figure 4.4: Biotoxins concentration above the safety limit.

represent 2/3 of the areas illustrated is not surprising due to the concept of the indicator species. In many production areas, it is only when contamination is detected in the mussel that the measurement of other species is preceded. Table 4.3 shows the disparity between mussels observations and the remaining most observed species.

Table 4.3: Species with the most data recorded.

Species	Mussel	Cockle	Pacific Oyster	Donax Clam	Surf Clam	Grooved Carpet Shell
Count	6064	2047	892	844	788	778

Table 4.4: Percentage of missing data by production area. Bold digits mean a variable was discarded due to too much missing data.

	Percentage of missing (%)								
	L1	L2	L3	L5b	L8	L9	RIAV1	RIAV2(M)	RIAV2(B)
DSP toxins	0.33	0.24	**0.62**	0.15	0.24	0.23	0.1	0.13	0.08
DSP phyto	0.08	0.06	0.03	**0.54**	0.35	0.05	0.04	0.05	0.05
SST	0	0	0	0	0	0	0	0	0
Chl-a	0	0	0	0	0	0	0	0	0
Max temp	0.11	0.07	0	0	0	0	0	0	0
Wind int	**0.68**	0.07	0.27	0	0	**1**	0.27	0.27	0.27
Wind dir	**0.95**	**0.52**	0.35	**0.95**	**0.95**	**1**	0.35	0.35	0.35
Rainfall	0.01	0.35	0.01	0.02	0.02	0.46	0.01	0.01	0.01

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. Table 4.4 shows the percentage of missing data per variable in each production zone. In addition to the features associated with the ASP and PSP toxins having already been removed, 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.

4.1.4 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 ($\hat{y}$), and the best method is the one that minimizes the root mean squared error (RMSE) denoted mathematically in Equation 4.1:

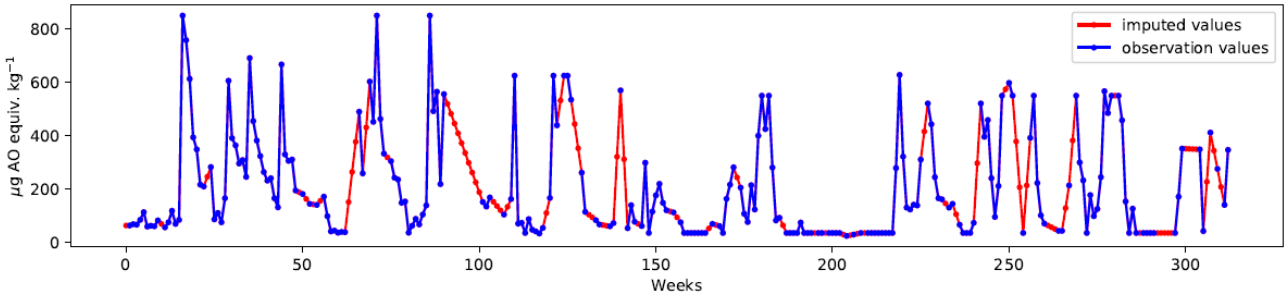
$$\text{RMSE} = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(y_i - \hat{y}_i)^2}, \tag{4.1}$$

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. For instance, the production area L2 obtained as imputation method: *linear*, *knn*, *linear*, *nocb*, *knn* and *nn* for the following variables, respectively: *DSP toxins*, *DSP phyto*, *Max temp*, *Wind int*, *Wind dir* and *Rainfall*. These results are expressed in Figure 4.5.

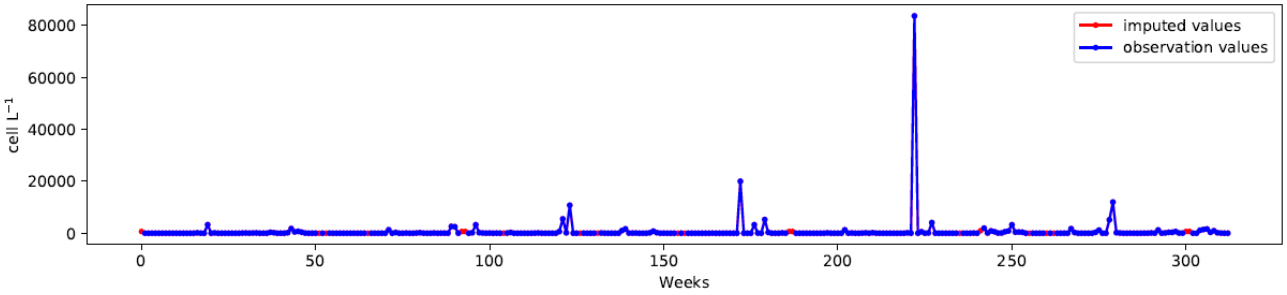
4.1.5 Data Transformation

4.1.5.1 Data Stationarity

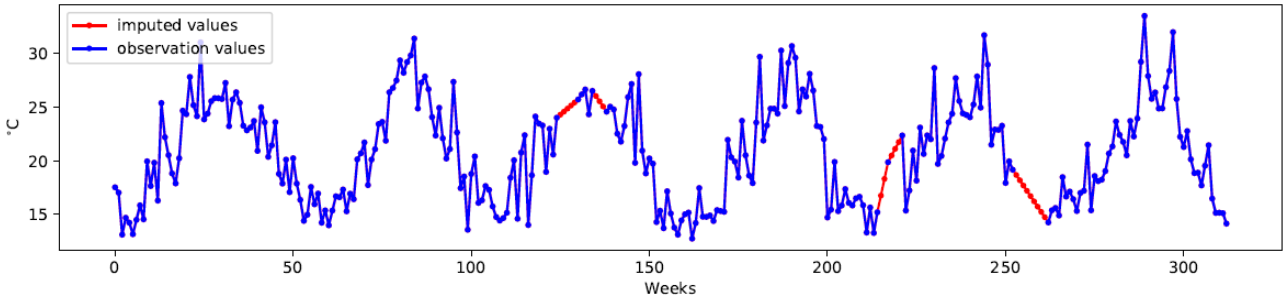
A stationary analysis was conducted to avoid the spurious correlations that non-stationary time series may introduce. For analysing stationarity, a combination of the augmented Dickey-Fuller (ADF) test [94] test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [95] was performed. This system can 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.



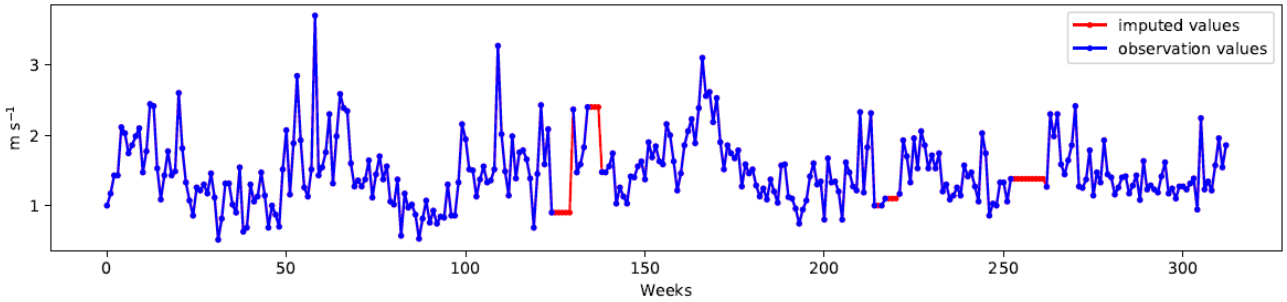
(a) Linear imputation of DSP toxins concentration



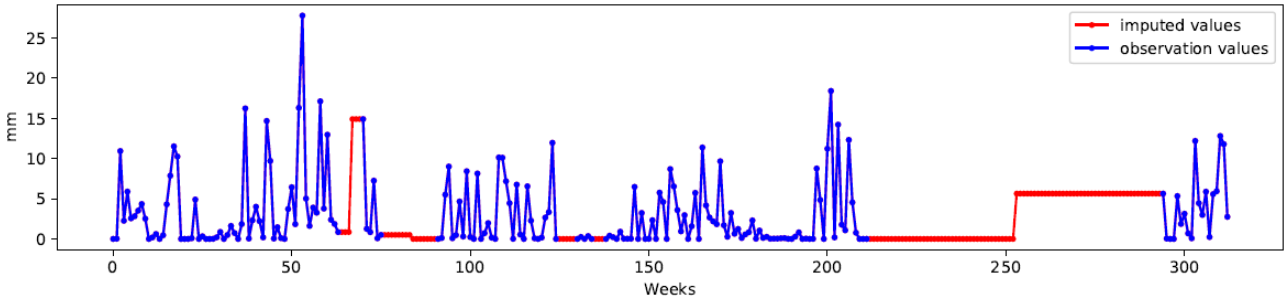
(b) Knn imputation of DSP toxins-producing phytoplankton cell counts.



(c) Linear imputation of maximum air temperature.



(d) Nocb imputation of mean wind speed.



(e) Nn imputation of rainfall.

Figure 4.5: L2 production area imputation techniques.

4.1.5.2 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. This technique resizes data values to a range between 0 and 1. Considering that X is the original data point, X_{max} is the maximum value and X_{min} is the minimum value in the dataset, the normalization value is given by:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.2}$$

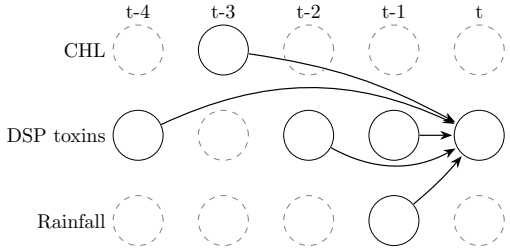
4.2 Results and Discussion

This section presents the cause-effect relationships determined by PCMCI+ using GPDC with $\alpha_{\mathrm{PC}} = 0.3$ between DSP toxins and the biological (DSP toxins-producing phytoplankton), oceanographic (SST and *chl-a*) and meteorological (maximum temperature, wind intensity, wind direction and rainfall) variables between 2015 and 2020. Thus, an analysis of potential predictors of contamination is explored for the three species mentioned, namely mussels, cockles and donax clams.

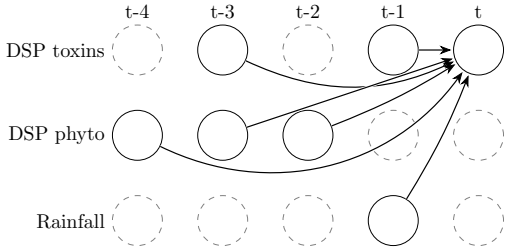
4.2.1 Mussel Analysis

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 4.1, the causal algorithm selected in Section 3 was used to illustrate in Figure 4.6 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 [96, 97].

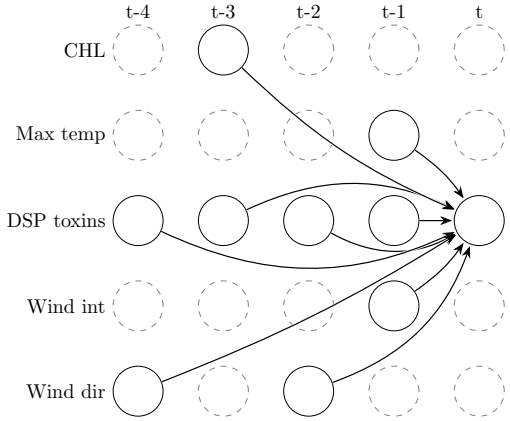
This approach alone does not allow us to isolate which predictors contribute most to mussel contamination above safety limits. Still, it does allow us to more generally infer which are the most important variables in the toxicity ecosystem and which temporal lag impacts mussel toxicity. By selecting production areas with more contamination cases, we are closer to achieving characteristics related to high toxicity. Thus, Figure 4.7 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 4.7b) and another that, instead of penalizing contaminants detected with different temporal lags, it benefits if the same predictors exist in different production areas (Figure 4.7a). Without discretizing the temporal lags, we can state that the DSP autocorrelation presents good evidence of being able to justify



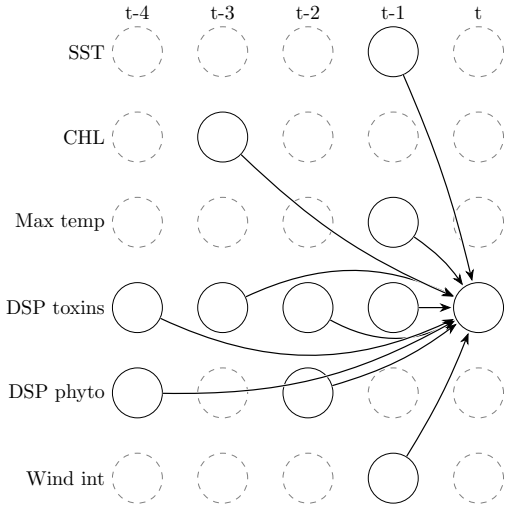
(a) L1 production area.



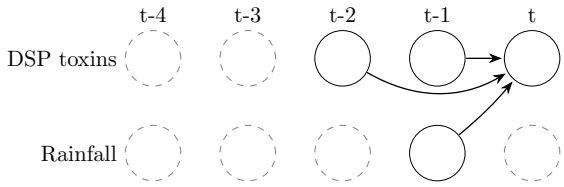
(b) L2 production area.



(c) RIAV1 production area.



(d) RIAV2(M) production area.

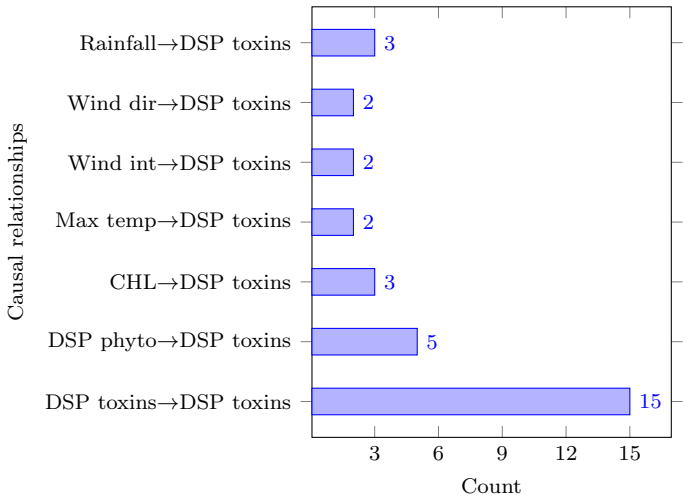


(e) L5b production area.

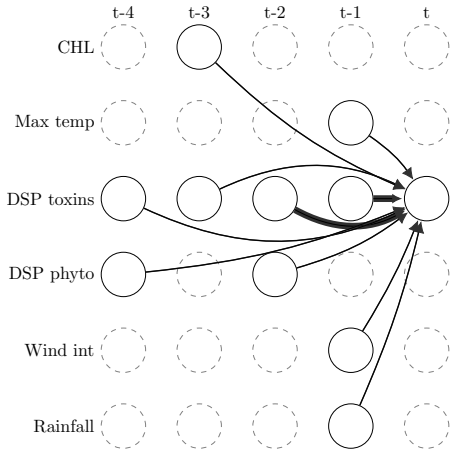
Figure 4.6: Causal relationships detected in contaminated mussels with a temporal lag up to four weeks.

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 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 4.6.

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 4.8b 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.



(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.

Figure 4.7: Aggregated results for mussels contamination with discretized lags (a) and without discretized lags (b).

DSP toxins-producing phytoplankton report the same importance as rainfall in the approach without discretized temporal lags (Figure 4.8a). The fact that phytoplankton does not appear in the discretized results only means that it was detected with three different lags in Figure 4.6.

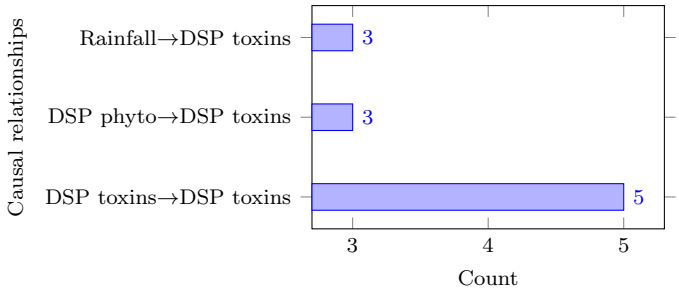
Figure 4.8d indicates that the DSP toxins present a very strong autocorrelation for coastal areas, highlighting 1-week and 2-week. Additionally, it was discovered that *chl-a*, maximum air temperature and wind intensity do not count as contaminants for coastal areas. Furthermore, rainfall does not count as a predictor of contamination for estuarine-lagoon areas.

4.2.1.1 Mussel Summer/Winter Analysis

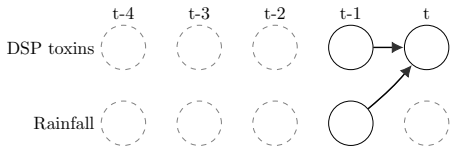
Finally, an analysis of contamination during summer and winter is carried out. The objective would be to understand if there is any hidden regime in the data during this period. To this end, the summer period was considered between 20/03 - 23/09, and the winter period between 01/01 - 20/03 and 23/09 - 31/12. A comparison will be conducted against Figure 4.7b, which illustrates the six years of contamination.

In the summer period, represented in Figure 4.9b, 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.

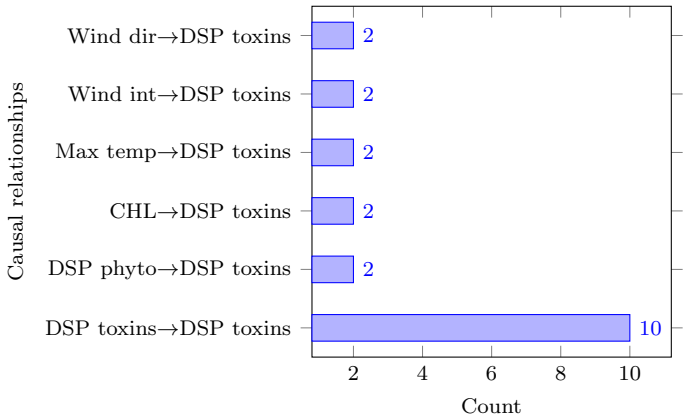
Concerning the winter period, *chl-a*, DSP toxin-producing phytoplankton, wind intensity and rainfall are no longer detected. Maximum air temperature and DSP autocorrelation manifest mostly as short-dependency causal links.



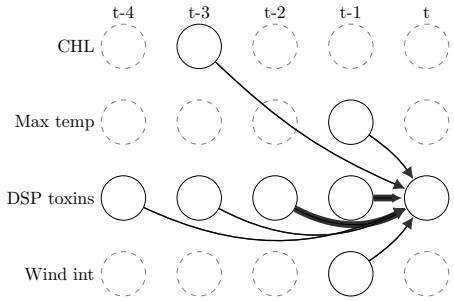
(a) Coastal zone without discretized temporal lags



(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.

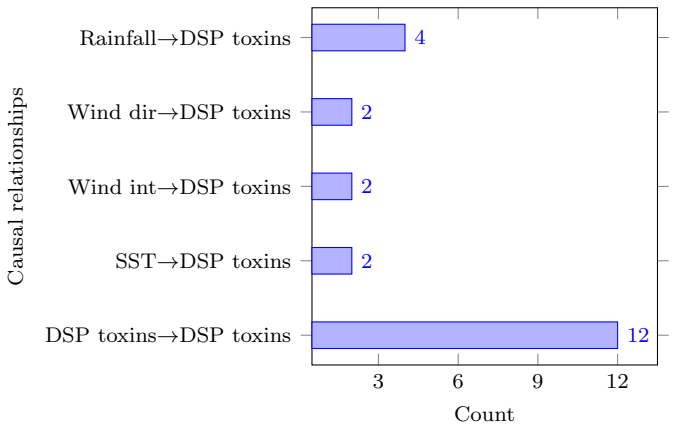
Figure 4.8: Aggregated results for coastal and estuarine-lagoon zones in mussels contamination with discretized lags (a) and without discretized lags (b).

4.2.2 Donax Clam Analysis

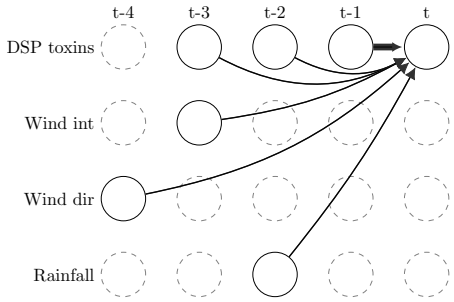
The donax clam *Donax trunculus* with the most cases of DSP toxins contamination are in the following production areas: L8 and L9. Figure 4.10 shows the causal relationships detected between contaminated donax clams and environmental variables, up to a time lag of 4 weeks. The L9 production zone has 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.

4.2.2.1 Donax Clam Summer/Winter analysis

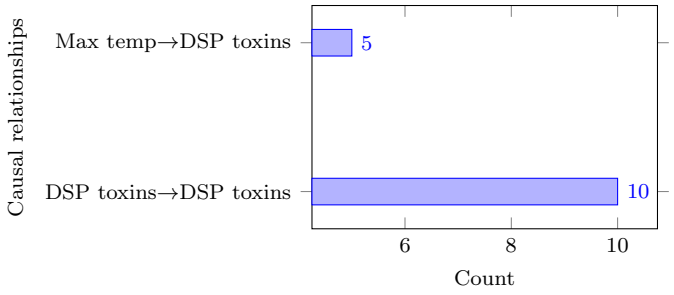
Regarding the analysis of donax clam contamination during summer and winter (Figure 4.11), only 1-week autocorrelation of DSP toxins was detected. This does not mean that there is no intrinsic regime for each production zone, but it does mean that there is no common regime between the two production areas. It would also not be expected that many contaminants would be found, since in Figure 4.10, only one DSP toxin contaminant was found (in addition to DSP toxins autocorrelation).



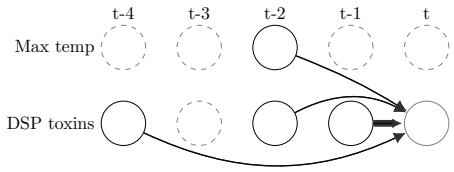
(a) Summer analysis without discretized temporal lags.



(b) Summer analysis with discretized temporal lags. There are thicker edges to record whether causal links were found in 2 or 3 (less intense) and 4 or 5 production areas (more intense).

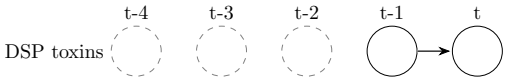


(c) Winter analysis without discretized temporal lags.

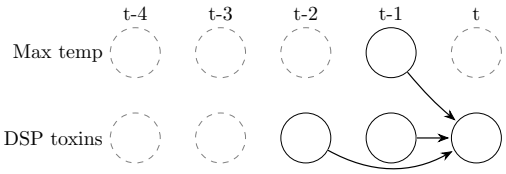


(d) Winter analysis with discretized temporal lags. There are thicker edges to record whether causal links were found in 2 or 3 (less intense) and 4 or 5 production areas (more intense).

Figure 4.9: Aggregated results in mussels contamination in summer and winter with discretized lags (a) and without discretized lags (b).



(a) L8 production area.

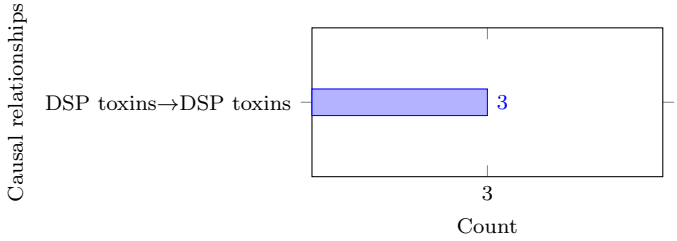


(b) L9 production area.

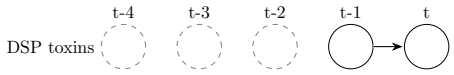
Figure 4.10: Causal relationships detected in contaminated donax clam with a temporal lag up to four weeks.

4.2.3 Cockle Analysis

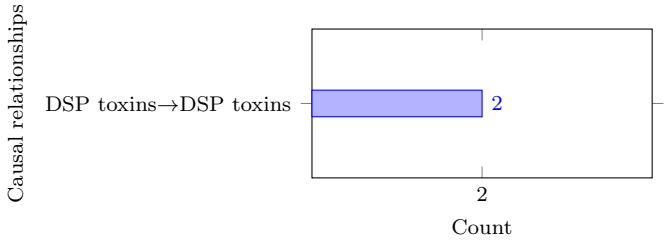
The cockle *Cerastoderma edule* with the highest number of cases of DSP toxins contamination is in the RIAV2(B) production area. This analysis will be the most restricted because it only concerns one production area, even in terms of validating results with adjacent areas. Figure 4.12 pictures the causal 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.



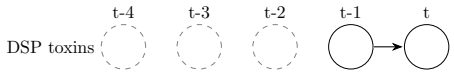
(a) Summer analysis without discretized temporal lags.



(b) Summer analysis with discretized temporal lags. Only causal links detected in both areas are reported.



(c) Winter analysis without discretized temporal lags.



(d) Winter analysis with discretized temporal lags. Only causal links detected in both areas are reported.

Figure 4.11: Aggregated results in donax clams contamination in summer and winter with discretized lags (a) and without discretized lags (b).

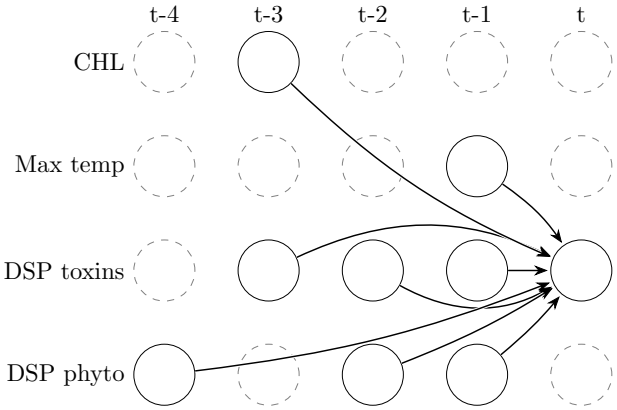
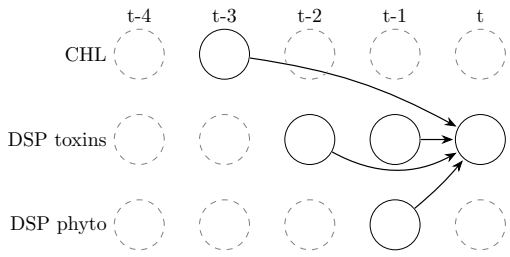


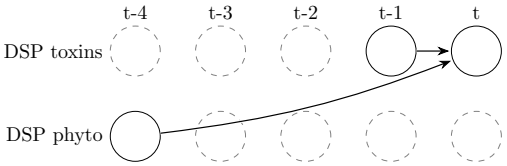
Figure 4.12: Causal relationships detected in contaminated cockles in the RIAV2(B) production area with a temporal lag up to four weeks.

4.2.3.1 Cockle Summer/Winter analysis

To conclude, Figure 4.13 shows cockle contamination during summer and winter. In both periods, 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 (Figure 4.13a), 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.



(a) Summer analysis with discretized temporal lags.



(b) Winter analysis with discretized temporal lags. Only causal links detected in both areas are reported.

Figure 4.13: Aggregated results in cockles contamination in summer and winter with discretized lags (a) and without discretized lags (b).

Chapter 5

Conclusions 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.

The temporary and unpredictable closure of shellfish production areas continues to pose a challenge. Robust monitoring programmes must be carried out in several harvesting areas to safeguard public health and minimise the economic losses from the local farmers. Identifying the contamination phenomenon and the impact of collected environmental variables can aid in constructing more precise forecasting systems to detect shellfish toxicity. Although forecasting methods 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 most significant impact on shellfish toxicity and HAB events. In this way, this thesis was designed to provide more insights and explainability of the variables most present in the role of contamination by using causality methods to infer causal relationships that govern the underlying structure of time series.

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 $\alpha_{\mathrm{PC}} = 0.3$ stood out.

Lastly, we examined the time series relating to our case study, i.e. the contamination of mollusc bivalves along the Portuguese coast. This analysis was divided into two phases: data preprocessing stages and the results obtained by the causal algorithm. The former is based on data modifications so that future research can be carried out more successfully. To summarise, of the 41 production areas

available, only eight were selected based on the concentration of DSP toxins. ASP and PSP toxins were excluded due to insufficient events above the regulatory limit. Production areas L1, L2, L5b, RIAV1 and RIAV2(M) were chosen to analyse mussel contamination, L8 and L9 to examine donax clam contamination and RIAV2(B) to explore cockle contamination.

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 to improve forecasting systems.

Bibliography

[1] https://www.statista.com/outlook/cmo/food/fish-seafood/fresh-seafood/portugal?currency=EUR.

[2] M. Mateus, J. Fernandes, M. Revilla, L. Ferrer, M. R. Villarreal, P. Miller, W. Schmidt, J. Maguire, A. Silva, and L. Pinto. Early warning systems for shellfish safety: The pivotal role of computational science. In J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, and P. M. Sloot, editors, *Computational Science – ICCS 2019*, pages 361–375, Cham, 2019. Springer International Publishing. ISBN 978-3-030-22747-0.

[3] S. Martino, F. Gianella, and K. Davidson. An approach for evaluating the economic impacts of harmful algal blooms: The effects of blooms of toxic dinophysis spp. on the productivity of scottish shellfish farms. *Harmful Algae*, 99:101912, 2020. ISSN 1568-9883. doi: https://doi.org/10.1016/j.hal.2020.101912. URL https://www.sciencedirect.com/science/article/pii/S1568988320301918.

[4] B. Dale, M. Edwards, and P. Reid. *Climate Change and Harmful Algal Blooms*, volume 189, pages 367–378. 01 2006. ISBN 978-3-540-32209-2. doi: 10.1007/978-3-540-32210-8_28.

[5] L. Avdelas, E. Avdic-Mravlje, A. C. Borges Marques, S. Cano, J. J. Capelle, N. Carvalho, M. Cozzolino, J. Dennis, T. Ellis, J. M. Fernández Polanco, J. Guillen, T. Lasner, V. Le Bihan, I. Llorente, A. Mol, S. Nicheva, R. Nielsen, H. van Oostenbrugge, S. Villasante, S. Visnic, K. Zhelev, and F. Asche. The decline of mussel aquaculture in the european union: causes, economic impacts and opportunities. *Reviews in Aquaculture*, 13(1):91–118, 2021. doi: https://doi.org/10.1111/raq.12465. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12465.

[6] J. Mardones, D. Holland, L. Anderson, L. B. Véronique, F. Gianella, A. Clement, K. Davidson, S. Sakamoto, T. Yoshida, and V. Trainer. *Estimating and Mitigating the Economic Costs of Harmful Algal Blooms on Commercial and Recreational Shellfish Harvesters*, pages 66–83. 11 2020. ISBN 978-1-927797-40-2.

[7] L. M. Grattan, S. Holobaugh, and J. G. Morris. Harmful algal blooms and public health. *Harmful Algae*, 57:2–8, 2016. ISSN 1568-9883. doi: https://doi.org/10.1016/j.hal.2016.05.003. URL https://www.sciencedirect.com/science/article/pii/S1568988316301664. Harmful Algal Blooms and Public Health.

[8] C. o. t. E. U. European Parliament. Regulation (ec) no 853/2004 of the european parliament and of

the council of 29 april 2004. *Official Journal of the European Union*. URL https://www.ipma.pt/pt/bivalves/docs/index.jsp.

[9] 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*, 8, 2021. ISSN 2296-7745. doi: 10.3389/fmars.2021.666583. URL https://www.frontiersin.org/articles/10.3389/fmars.2021.666583.

[10] R. Salas and D. Clarke. Review of dsp toxicity in ireland: Long-term trend impacts, biodiversity and toxin profiles from a monitoring perspective. *Toxins*, 11(2), 2019. ISSN 2072-6651. doi: 10.3390/toxins11020061. URL https://www.mdpi.com/2072-6651/11/2/61.

[11] K. Davidson, C. Whyte, D. Aleynik, A. Dale, A. Kurekin, S. Mcneill, P. Miller, M. Porter, and R. Saxon. Habreports:: Online early warning of harmful algal and biotoxin risk for the shellfish and finfish aquaculture industries. *Frontiers in Marine Science*, 8, Apr. 2021. ISSN 2296-7745. doi: 10.3389/fmars.2021.631732. © 2021 Davidson, Whyte, Aleynik, Dale, Kurekin, Gontarek, McNeill, Miller, Porter, Saxon and Swan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

[12] A. Silva, L. Pinto, S. Rodrigues, H. de Pablo, M. Santos, T. Moita, and M. Mateus. A hab warning system for shellfish harvesting in portugal. *Harmful Algae*, 53:33–39, 2016. ISSN 1568-9883. doi: https://doi.org/10.1016/j.hal.2015.11.017. URL https://www.sciencedirect.com/science/article/pii/S1568988315001730. Applied Simulations and Integrated Modelling for the Understanding of Toxic and Harmful Algal Blooms (ASIMUTH).

[13] J. H. Lee, Y. Huang, M. Dickman, and A. Jayawardena. Neural network modelling of coastal algal blooms. *Ecological Modelling*, 159(2):179–201, 2003. ISSN 0304-3800. doi: https://doi.org/10.1016/S0304-3800(02)00281-8. URL https://www.sciencedirect.com/science/article/pii/S0304380002002818.

[14] F. N. Yussof, N. Maan, and M. N. Md Reba. Lstm networks to improve the prediction of harmful algal blooms in the west coast of sabah. *International Journal of Environmental Research and Public Health*, 18(14), 2021. ISSN 1660-4601. doi: 10.3390/ijerph18147650. URL https://www.mdpi.com/1660-4601/18/14/7650.

[15] 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*, 9(3), 2021. ISSN 2077-1312. doi: 10.3390/jmse9030283. URL https://www.mdpi.com/2077-1312/9/3/283.

[16] R. C. Cruz, P. R. Costa, L. Krippahl, and M. B. Lopes. Forecasting biotoxin contamination in mussels across production areas of the portuguese coast with artificial neural networks. *Knowledge-Based*

Systems, 257:109895, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109895. URL https://www.sciencedirect.com/science/article/pii/S0950705122009881.

[17] A. da Silva Pereira. Time series analysis and forecasting of shellfish contamination and safety. MEIC Thesis in Information Systems and Computer Engineering, IST University of Lisbon, Alameda, Lisboa, Portugal, October 2021.

[18] M. S. Madeira. Multivariate time-series modeling of shellfish contamination with dynamic bayesian networks. MECD Thesis in Computer Science and Engineering, IST University of Lisbon, Alameda, Lisboa, Portugal, June 2022.

[19] A. Patrício, M. B. Lopes, P. R. Costa, R. S. Costa, R. Henriques, and S. Vinga. Time-lagged correlation analysis of shellfish toxicity reveals predictive links to adjacent areas, species, and environmental conditions. *Toxins*, 14(10), 2022. ISSN 2072-6651. doi: 10.3390/toxins14100679. URL https://www.mdpi.com/2072-6651/14/10/679.

[20] A. C. Braga, S. M. Rodrigues, H. M. Lourenço, P. R. Costa, and S. Pedro. Bivalve shellfish safety in portugal: Variability of faecal levels, metal contaminants and marine biotoxins during the last decade (2011-2020). *Toxins*, 15(2), 2023. ISSN 2072-6651. doi: 10.3390/toxins15020091. URL https://www.mdpi.com/2072-6651/15/2/91.

[21] M. J. Botelho, C. Vale, and J. G. Ferreira. Seasonal and multi-annual trends of bivalve toxicity by psts in portuguese marine waters. *Science of The Total Environment*, 664:1095–1106, 2019. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2019.01.314. URL https://www.sciencedirect.com/science/article/pii/S0048969719303572.

[22] P. Vale, S. Gomes, M.-j. Botelho, and S. Rodrigues. Monitorizaçao de psp na costa portuguesa através de espécies-indicadoras. 01 2007.

[23] P. Vale, M. J. Botelho, S. M. Rodrigues, S. S. Gomes, and M. A. de M. Sampayo. Two decades of marine biotoxin monitoring in bivalves from portugal (1986–2006): A review of exposure assessment. *Harmful Algae*, 7(1):11–25, 2008. ISSN 1568-9883. doi: https://doi.org/10.1016/j.hal.2007.05.002. URL https://www.sciencedirect.com/science/article/pii/S156898830700087X.

[24] A. Papana. Connectivity analysis for multivariate time series: Correlation vs. causality. *Entropy*, 23(12), 2021. ISSN 1099-4300. doi: 10.3390/e23121570. URL https://www.mdpi.com/1099-4300/23/12/1570.

[25] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1), June 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10105-3.

[26] M. Kretschmer, D. Coumou, J. Donges, and J. Runge. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, 29:160303130523003, 03 2016. doi: 10.1175/JCLI-D-15-0654.1.

[27] J. A. McGowan, E. R. Deyle, H. Ye, M. L. Carter, C. T. Perretti, K. D. Seger, A. de Verneil, and G. Sugihara. Predicting coastal algal blooms in southern california. *Ecology*, 98(5):1419–1433, 2017. doi: https://doi.org/10.1002/ecy.1804. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.1804.

[28] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

[29] M. Nauta, D. Bucur, and C. Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL https://www.mdpi.com/2504-4990/1/1/19.

[30] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996. URL https://www.science.org/doi/abs/10.1126/sciadv.aau4996. Code available on, https://github.com/jakobrunge/tigramite.

[31] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. URL http://jmlr.org/papers/v11/hyvarinen10a.html.

[32] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, and B. Aragam. Dynotears: Structure learning from time-series data. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1595–1605. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html. Code available on, https://github.com/quantumblacklabs/causalnex.

[33] Copernicus. Available on,https://www.copernicus.eu/.

[34] IPMA. Available on, https://www.ipma.pt/pt/bivalves/.

[35] B. Neal. Introduction to causal inference. 2020.

[36] P. Bhandari. Correlation vs. causation: Difference, designs and examples. 2021.

[37] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, 53(4), jul 2020. ISSN 0360-0300. doi: 10.1145/3397269. URL https://doi.org/10.1145/3397269.

[38] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524.

[39] M. J. Vowels, N. C. Camgoz, and R. Bowden. D'ya like dags? a survey on structure learning and causal discovery, 2021. URL https://arxiv.org/abs/2103.02582.

[40] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, 2022. doi: https://doi.org/10.1002/widm.1449. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1449.

[41] J. L. Monteiro, S. Vinga, and A. M. Carvalho. Polynomial-time algorithm for learning optimal tree-augmented dynamic Bayesian networks. In M. Meila and T. Heskes, editors, *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 622–631. AUAI Press, 2015.

[42] A. M. Carvalho, P. Adão, and P. Mateus. Hybrid learning of bayesian multinets for binary classification. *Pattern Recognit.*, 47(10):3438–3450, 2014.

[43] B. Hannisdal and L. H. Liow. Causality from palaeontological time series. *Palaeontology*, 61(4): 495–509, 2018. doi: https://doi.org/10.1111/pala.12370. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pala.12370.

[44] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 07 2018. ISSN 1054-1500. doi: 10.1063/1.5025050. URL https://doi.org/10.1063/1.5025050.

[45] C. K. Assaad, E. Devijver, and E. Gaussier. Survey and evaluation of causal discovery methods for time series. *J. Artif. Int. Res.*, 73, may 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://doi.org/10.1613/jair.1.13428.

[46] M. Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1997): 20110613, 2013. doi: 10.1098/rsta.2011.0613. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0613.

[47] Y. Hmamouche, A. Casali, and L. Lakhal. A causality-based feature selection approach for multivariate time series forecasting. 07 2017.

[48] https://commons.wikimedia.org/wiki/File:GrangerCausalityIllustration.svg.

[49] Y. Sun, J. Li, J. Liu, C. Chow, B.-Y. Sun, and R. Wang. Using causal discovery for feature selection in multivariate numerical time series. *Machine Learning*, 101:377–395, 2015.

[50] R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, and H. Liu. Causal inference for time series analysis: Problems, methods and evaluation, 2021. URL https://arxiv.org/abs/2102.05829.

[51] S. Löwe, D. Madras, R. S. Zemel, and M. Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. *CoRR*, abs/2006.10833, 2020. URL https://arxiv.org/abs/2006.10833.

[52] W. Peng. DLI: A Deep Learning-Based Granger Causality Inference. *Complexity*, 2020:1–6, June 2020. doi: 10.1155/2020/5960171. URL https://ideas.repec.org/a/hin/complx/5960171.html.

[53] Y. Cheng, R. Yang, T. Xiao, Z. Li, J. Suo, K. He, and Q. Dai. Cuts: Neural causal discovery from irregular time-series data, 2023.

[54] M. Nauta, D. Bucur, and C. Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL https://www.mdpi.com/2504-4990/1/1/19.

[55] U. Hasan, E. Hossain, and M. O. Gani. A survey on causal discovery methods for i.i.d. and time series data. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=YdMrdhGx9y. Survey Certification.

[56] P. Spirtes. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review - SOC SCI COMPUT REV*, 9:62–72, 04 1991. doi: 10.1177/089443939100900106.

[57] F. R. d. Almeida. Causal discovery from time series data. European master of medical technology and healthcare business, Politecnico do Porto - Escola Superior de Saude, Alameda, Lisboa, Portugal, Setember 2022.

[58] J. Runge, J. Heitzig, N. Marwan, and J. Kurths. Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E*, 86(6), dec 2012. doi: 10.1103/physreve.86.061121. URL https://doi.org/10.1103%2Fphysreve.86.061121.

[59] B. Pompe and J. Runge. Momentary information transfer as a coupling measure of time series. *Phys. Rev. E*, 83:051122, May 2011. doi: 10.1103/PhysRevE.83.051122. URL https://link.aps.org/doi/10.1103/PhysRevE.83.051122.

[60] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014. URL http://jmlr.org/papers/v15/colombo14a.html.

[61] S. Peterson. Comparison of lasso granger and pcmci for causal feature selection in multivariate time series, 2022. URL http://hdl.handle.net/10150/665005.

[62] J. Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, 2020. URL https://arxiv.org/abs/2003.03685.

[63] S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL http://jmlr.org/papers/v7/shimizu06a.html.

[64] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL http://jmlr.org/papers/v12/shimizu11a.html.

[65] R. Howard and L. Kunze. Evaluating temporal observation-based causal discovery techniques applied to road driver behaviour, 2023.

[66] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

[67] B. D. Deaton. Effects of the swiss franc/euro exchange rate floor on the calibration of probability forecasts. *Forecasting*, 1(1):3–25, 2019. ISSN 2571-9394. doi: 10.3390/forecast1010002. URL https://www.mdpi.com/2571-9394/1/1/2.

[68] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018.

[69] M. B. Wang and M. Kronovet. Time-varying dags with notears final report. 2020. URL https://api.semanticscholar.org/CorpusID:231775505.

[70] C. Amornbunchornvej, E. Zheleva, and T. Berger-Wolf. Variable-lag granger causality for time series analysis. pages 21–30, 10 2019. doi: 10.1109/DSAA.2019.00016.

[71] A. E. Yuan and W. Shou. Data-driven causal analysis of observational biological time series. *eLife*, 11:e72518, aug 2022. ISSN 2050-084X. doi: 10.7554/eLife.72518. URL https://doi.org/10.7554/eLife.72518.

[72] A. Attanasio, A. Pasini, and U. Triacca. A contribution to attribution of recent global warming by out-of-sample granger causality analysis. *Atmospheric Science Letters*, 13(1):67–72, 2012. doi: https://doi.org/10.1002/asl.365. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.365.

[73] E. Kodra, S. Chatterjee, and A. Ganguly. Exploring granger causality between global average observed time series of carbon dioxide and temperature. *Theoretical and Applied Climatology*, 104: 325–335, 06 2011. doi: 10.1007/s00704-010-0342-3.

[74] C. Papagiannopoulou, S. Decubber, D. G. Miralles, M. Demuzere, N. E. C. Verhoest, and W. Waegeman. Analyzing granger causality in climate data with time series classification methods. In Y. Altun,

K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Žitnik, M. Ceci, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 15–26, Cham, 2017. Springer International Publishing. ISBN 978-3-319-71273-4.

[75] C. Papagiannopoulou, D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman. A non-linear granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960, 2017. doi: 10.5194/gmd-10-1945-2017. URL https://gmd.copernicus.org/articles/10/1945/2017/.

[76] Y. Huang, M. Kleindessner, A. Munishkin, D. Varshney, P. Guo, and J. Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in Big Data*, 4, 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.642182. URL https://www.frontiersin.org/articles/10.3389/fdata.2021.642182.

[77] C. Krich, J. Runge, D. G. Miralles, M. Migliavacca, O. Perez-Priego, T. El-Madany, A. Carrara, and M. D. Mahecha. Estimating causal networks in biosphere–atmosphere interaction with the pcmci approach. *Biogeosciences*, 17(4):1033–1061, 2020. doi: 10.5194/bg-17-1033-2020. URL https://bg.copernicus.org/articles/17/1033/2020/.

[78] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, 29(11):4069 – 4081, 2016. doi: https://doi.org/10.1175/JCLI-D-15-0654.1. URL https://journals.ametsoc.org/view/journals/clim/29/11/jcli-d-15-0654.1.xml.

[79] C. Krich, M. D. Mahecha, M. Migliavacca, M. G. D. Kauwe, A. Griebel, J. Runge, and D. G. Miralles. Decoupling between ecosystem photosynthesis and transpiration: a last resort against overheating. *Environmental Research Letters*, 17(4):044013, mar 2022. doi: 10.1088/1748-9326/ac583e. URL https://dx.doi.org/10.1088/1748-9326/ac583e.

[80] J. Runge, V. Petoukhov, and J. Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27(2):720 – 739, 2014. doi: https://doi.org/10.1175/JCLI-D-13-00159.1. URL https://journals.ametsoc.org/view/journals/clim/27/2/jcli-d-13-00159.1.xml.

[81] G. Menegozzo, D. Dall'Alba, and P. Fiorini. Causal interaction modeling on ultra-processed food manufacturing. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 200–205, 2020. doi: 10.1109/CASE48305.2020.9216973.

[82] D. M. Álvarez and G. Poveda. Spatiotemporal dynamics of ndvi, soil moisture and enso in tropical south america. *Remote Sensing*, 14(11), 2022. ISSN 2072-4292. doi: 10.3390/rs14112521. URL https://www.mdpi.com/2072-4292/14/11/2521.

[83] T. Ikeuchi, M. Ide, Y. Zeng, T. N. Maeda, and S. Shimizu. Python package for causal discovery based on lingam. *Journal of Machine Learning Research*, 24(14):1–8, 2023. URL http://jmlr.org/papers/v24/21-0321.html. Code available on, https://github.com/cdt15/lingam.

[84] H. Ohmura. The connection between stock market prices and political support: evidence from Japan. *Applied Economics Letters*, 29(1):1–7, January 2022. doi: 10.1080/13504851.2020.185. URL https://ideas.repec.org/a/taf/apeclt/v29y2022i1p1-7.html.

[85] A. Einizade and S. H. Sardouie. Estimation of a causal directed acyclic graph process using non-gaussianity, 2022.

[86] M. H. Ferdous, U. Hasan, and M. O. Gani. Ecdans: Efficient temporal causal discovery from autocorrelated and non-stationary data (student abstract). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i13.26964. URL https://doi.org/10.1609/aaai.v37i13.26964.

[87] X. Yang, C. Zhang, and B. Zheng. Segment-wise time-varying dynamic bayesian network with graph regularization. *ACM Transactions on Knowledge Discovery from Data*, 16, 05 2022. doi: 10.1145/3522589.

[88] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54(2):875–891, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.08.063. URL https://www.sciencedirect.com/science/article/pii/S1053811910011602.

[89] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL http://jmlr.org/papers/v17/14-518.html.

[90] A. Mudadu, A. Bazzoni, R. Bazzardi, G. Lorenzoni, B. Soro, N. Bardino, I. Arras, G. Sanna, B. Vodret, and S. Virgilio. Influence of seasonality on the presence of okadaic acid associated with dinophysis species: A four-year study in sardinia (italy). *Italian Journal of Food Safety*, 10, 03 2021. doi: 10.4081/ijfs.2021.8947.

[91] P. Vale. Two simple models for accounting mussel contamination with diarrhoetic shellfish poisoning toxins at aveiro lagoon: Control by rainfall and atmospheric forcing. *Estuarine, Coastal and Shelf Science*, 98:94–100, 2012. ISSN 0272-7714. doi: https://doi.org/10.1016/j.ecss.2011.12.007. URL https://www.sciencedirect.com/science/article/pii/S027277141100521X.

[92] T. C.-H. Lee, F. L.-Y. Fong, K.-C. Ho, and F. W.-F. Lee. The mechanism of diarrhetic shellfish poisoning toxin production in prorocentrum spp.: Physiological and molecular perspectives. *Toxins*, 8 (10), 2016. ISSN 2072-6651. doi: 10.3390/toxins8100272. URL https://www.mdpi.com/2072-6651/8/10/272.

[93] W. Schmidt, H. Evers-King, C. Campos, D. Jones, P. Miller, K. Davidson, and J. Shutler. A generic approach for the development of short-term predictions of e. coli and biotoxins in shellfish. *Aquaculture Environment Interactions*, 10, 04 2018. doi: 10.3354/aei00265.

[94] D. Dickey and W. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74, 06 1979. doi: 10.2307/2286348.

[95] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. 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(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y.

[96] B. Reguera, P. Riobó, F. Rodríguez, P. Díaz, G. Pizarro, B. Paz, J. Franco, and J. Blanco. Dinophysis toxins: Causative organisms, distribution and fate in shellfish. *Marine drugs*, 12:394–461, 01 2014. doi: 10.3390/md12010394.

[97] A. C. Braga, R. Marçal, A. Marques, S. Guilherme, Óscar Vilariño, J. M. L. Martins, A. Gago-Martínez, P. R. Costa, and M. Pacheco. Invasive clams (ruditapes philippinarum) are better equipped to deal with harmful algal blooms toxins than native species (r. decussatus): evidence of species-specific toxicokinetics and dna vulnerability. *Science of The Total Environment*, 767:144887, 2021. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2020.144887. URL https://www.sciencedirect.com/science/article/pii/S0048969720384205.

Appendix A

Pseudo-codes

The pseudo-codes of the models introduced in Section 2.5 are presented below:

A.1 Pairwise Granger Causality pseudo-code

Algorithm 1 PWGC

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags
  Form an empty graph $\mathscr{G}$ with d nodes V
  Standardize the data and check if it is covariance stationary
  Find the optimal lag value $\tau \in \{1, \ldots, \tau_{\max}\}$
  for $\mathscr{X}^q \in V$ **do**
    Fit Mres: $(\mathscr{X}^q_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
    for $\mathscr{X}^p \in V \setminus \{\mathscr{X}^q\}$ **do**
      Fit Mfull: $(\mathscr{X}^p_{t-i}, \mathscr{X}^q_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
      z = test to compare (Mfull) and (Mres)
      if $z < \alpha$ **then** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
  Return the Summary DiGraph $\mathscr{G}$

Figure A.1: Pseudo-code for the Pairwise Granger Causality. Adapted from [45].

A.2 Multivariate Granger Causality pseudo-code

Algorithm 2 MVGC

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags
  Form an empty graph $\mathscr{G}$ with d nodes V
  Standardize the data and check if it is covariance stationary
  Find the optimal lag value $\tau \in \{1, \ldots, \tau_{\max}\}$
  for $\mathscr{X}^q \in V$ **do**
    Fit mvMfull: $(\mathscr{X}_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
    for $\mathscr{X}^p \in V \setminus \{\mathscr{X}^q\}$ **do**
      Fit mvMres: $(\mathscr{X}_{t-i} \setminus \{\mathscr{X}^p_{t-i}\})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
      z = test to compare (mvMfull) and (mvMres)
      if $z < \alpha$ **then** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
  Return the Summary DiGraph $\mathscr{G}$

Figure A.2: Pseudo-code for the Multivariate Granger Causality. Adapted from [45].

A.3 TCDF pseudo-code

Algorithm 3 TCDF

Require: $\mathscr{X}$ a d-dimensional time series of length T, number of hidden layers L, kernel size K, dilation coefficient c, number of epochs, loss function and learning rate
  $\tau_{max} = 1 + (K-1)\sum_{l=0}^{L} c^l$
  Form an empty graph $\mathscr{G}$ with $d\tau_{max}$ nodes V
  for $q \in \{1, \cdots, d\}$ **do**
    Fit $N_q : (\mathscr{X}_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}_t^q$
    Compute the attention scores a_q and the kernel weights W_q
    Sort the attention scores a_q into b with decreasing order
    Compute the biggest attention score s_q associated to the largest gap in b
    for $p \in \{1, \cdots, d\}$ **do**
      if $\sigma(a_{q,p}) > s_q$ **then**
        $i = \text{argmax}(W_{q,p,.})$
        Add edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ to $\mathscr{G}$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** add edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ to $\mathscr{G}$
    for $\mathscr{X}_{t-i}^p \in \text{Par}(\mathscr{X}_t^q, \mathscr{G})$ **do**
      Compute the loss of N_q on $\mathscr{X}$ where $\mathscr{X}_{t-i}^p$ is permuted
      if the loss increases significantly **then**
        Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ from $\mathscr{G}$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$
  Return the Window MAG $\mathscr{G}$

Figure A.3: Pseudo-code for the TCDF algorithm. Adapted from [45].

A.4 PCMCI pseudo-code

Algorithm 4 PCMCI

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
  Form an oriented graph $\mathscr{G}$ with $d\tau_{max}$ nodes V such that $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ for all $\mathscr{X}_{t-i}^p, \mathscr{X}_t^q \in V, i \in \{1, \cdots, \tau_{max}\}$
  for $\mathscr{X}_t^q \in V$ **do**
    $n = 0$
    while $\text{card}(\text{Par}(\mathscr{X}_t^q, \mathscr{G})) \geq n+1$ **do**
      for $\mathscr{X}_{t-i}^p \in \text{Par}(\mathscr{X}_t^q, \mathscr{G})$ s.t. $\text{card}(\text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \mathscr{X}_{t-i}^p) = n$ **do**
        $\mathscr{X}_{\mathbf{I}}^{\mathbf{R}} =$ first n variables of $\text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{\mathscr{X}_{t-i}^p\}$
        Compute $y_{q,p}$ the statistics that corresponds to the test $\mathscr{X}_{t-i}^p \perp\!\!\!\perp \mathscr{X}_t^q \mid \mathscr{X}_{\mathbf{I}}^{\mathbf{R}}$ and its p-value z
        if $z > \alpha$ **then**
          Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ from $\mathscr{G}$
          for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$
      Sort $\text{Par}(X_t^q, \mathscr{G})$ by decreasing order of the statistics $(y_{q,p})_p$
      $n = n+1$
  for $\mathscr{X}_t^q \in V$ **do**
    for $\mathscr{X}_{t-i}^p \in \text{Par}(X_t^q, \mathscr{G})$ s.t. $\text{card}(\text{Par}(\mathscr{X}_{t-i}^p, \mathscr{G})) > 0$ **do**
      Compute z the p-value that corresponds to the test $\mathscr{X}_t^q \perp\!\!\!\perp \mathscr{X}_{t-i}^p \mid \text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{\mathscr{X}_{t-i}^p\} \cup \text{Par}(\mathscr{X}_{t-i}^p, \mathscr{G})$
      if $z > \alpha$ **then**
        Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$
  Return the window DAG $\mathscr{G}$

Figure A.4: Pseudo-code for the PCMCI algorithm. Adapted from [45].

A.5 VAR-LiNGAM pseudo-code

Algorithm 5 VarLiNGAM

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
Form an empty graph $\mathscr{G}$ with $d\tau$ nodes V
Find the optimal lag $\tau \in \{1, \cdots, \tau_{max}\}$
Fit (VAR): $\mathscr{X} \mapsto \mathscr{X}$
Compute $(\mathbf{M}_\tau)_{1 \leq \tau \leq \tau_{\max}}$ the coefficients of (VAR) and $\xi =$ its residuals
$\mathbf{S} = \{1, \cdots, d\}$
while card$(\mathbf{S}) > 1$ **do**
  for $p \in \mathbf{S}$ **do**
    for $q \in \mathbf{S} \setminus \{p\}$ **do**
      Fit least squares regressions: $\xi^p \mapsto \xi^q$ and compute its residuals $\varepsilon^{p,q}$
    Compute y_p the statistics that corresponds to the test $\varepsilon^{p,\cdot} \perp\!\!\!\perp \xi^p$
  $p^* = \text{argmin}_{q \in \mathbf{S}} y_p$
  $\mathbf{S} = \mathbf{S} \setminus \{p^*\}$
  for $\mathscr{X}^q \in \mathscr{X}^{\mathbf{S}}$ **for** $i \in \{0, \cdots, \tau\}$ **do** add edge $\mathscr{X}^{p^*}_{t-i} \to \mathscr{X}^q_t$ to $\mathscr{G}$
Construct a strictly lower triangular matrix $\mathbf{A}_0$ by following the order in $\mathscr{G}$, and estimate the connection strengths $[\mathbf{A}_0]_{i,j}$ by using some conventional covariance-based.
for $i \in \{1, \cdots, \tau\}$ **do**
  $\mathbf{A}_i = (I - \mathbf{A}_0)\mathbf{M}_i$
Apply Adaptive Lasso on $\mathbf{A}$
for $i \in \{0, \cdots, \tau\}$ **do**
  for $q \in \{1, \cdots, d\}$ **do**
    for $p \in \{1, \cdots, d\}$ **do**
      if $[\mathbf{A}_i]_{p,q} = 0$ **then** remove edge $\mathscr{X}^p_{t-i} \to \mathscr{X}^q_t$ from $\mathscr{G}$
Return the window DAG $\mathscr{G}$

Figure A.5: Pseudo-code for the VAR-LiNGAM algorithm. Adapted from [45].

A.6 DYNOTEARS pseudo-code

Algorithm 6 DYNOTEARS

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, λ_A, λ_W, α
$W, A = \min_{W,A} f(W, A)$ from (Score)
for $w_{pq} \in W$ **do**
  if $w_{pq} \geq \alpha$ **then**
    Add $X^p_t \to X^q_t$ to $\mathscr{G}$
    for $(\mathscr{X}^p_j, \mathscr{X}^q_j) \in \text{Hom}(\mathscr{X}^p_t, \mathscr{X}^q_t, \mathscr{G})$ **do** add edge $\mathscr{X}^p_j \to \mathscr{X}^q_j$ to $\mathscr{G}$
for $i \in \{1, \cdots, \tau_{\max}\}$ **do**
  for $a_{pq} \in A_i$ **do**
    if $a_{pq} \geq \alpha$ **then**
      Add $X^p_{t-i} \to X^q_t$ to $\mathscr{G}$
      for $(\mathscr{X}^p_{j-i}, \mathscr{X}^q_j) \in \text{Hom}(\mathscr{X}^p_{t-i}, \mathscr{X}^q_t, \mathscr{G})$ **do** add edge $\mathscr{X}^p_{j-i} \to \mathscr{X}^q_j$ to $\mathscr{G}$
Return window DAG $\mathscr{G}$

Figure A.6: Pseudo-code for the DYNOTEARS algorithm. Adapted from [45].