Imperial College London

MEng Individual Project

Imperial College London

Department of Mathematics and Department of Computing

A Complex Systems Approach to Ecology: from Time-series Population Data to Food-Webs

Author:
Nikolay Smirnov

Supervisor:
Dr. Thibault Bertrand

Second Marker:
Prof. Henrik Jeldtoft Jensen

June 2020

Abstract

The Guadalquivir River estuary in the Gulf of Cadiz of Spain holds a pivotal role as the main feeding and nursery area of many juvenile species that are critical for the local economy and biodiversity conservation. We aim to disentangle the complex interactions that compose this dynamical system. We first provide a network-focused analysis of the food-web, finding that it exhibits small-world characteristics but does not appear scale-free. We apply a state-of-the-art causality detection method, known as convergent cross mapping, to predict predator-prey interactions from a large time-series dataset, for the dominant species in the ecosystem. We incorporate a phase-lock twin surrogate test to reduce the false positive predictions. Using this method we examine the relative impacts of environmental conditions on species and study the interaction of European eels, a species with a complex migratory pattern, with the ecosystem. This provides a practical demonstration of a recently developed technique on a large and intricate ecosystem.

Acknowledgements

I am incredibly grateful to my supervisor Dr. Thibault Bertrand for his unrelenting guidance and support, without which this project would not have been possible. I would like to also thank Dr. Cesar Vilas from the IFAPA, Spain, for his collaboration and the invaluable data he provided. Finally, I'd like to thank my friends and parents for their continuous support throughout my degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ecology is the study of organism relationships to one another and their physical surroundings. A core archetype of ecology is the study of food-webs - investigating the interactions between organisms and the transfer of energy by consumption. Food-webs can be constructed from combinations of food-chains, obtained through experimental observation and mathematical models of consumer-resource system dynamics. This is an active area of investigation, with most studies focused on the network properties of food-webs and no general model for creating food-webs[1]. For some systems, species population data may be simpler to measure than all consumer interactions, and a model that can represent the food-web using species population data will provide an improved intuition into the ecosystem's dynamics.

The Guadalquivir River estuary, Spain, provides an interesting ecosystem for ecological research. It is the main feeding area and nursery for the early life stages of many species that are key to the local economy, influencing many sectors, the main one being tourism. Some of these species, such as the European eel, sea bass and river herring, are at risk of extinction and are targets of European protection and recovery programs, making the area also important for biodiversity conservation efforts[2]. For these reasons, the area has attracted research funding from Spanish and European agencies, leading to the Long Term Ecological Research (LTER) program conducting work there. Over 18 consecutive years, the estuary has been surveyed in multiple locations for species abundances and many environmental conditions, providing a rich dataset[2]. In addition to this, a food-web of the ecosystem has been created from a combination of studies[2]. The estuary is a very complex open system, with many inputs, outputs, and varying environmental factors influencing the system, but we shall attempt to model it using this data.

Due to the dynamic nature of ecosystems, the abundance of one species has the ability to alter the abundance of other species and the state of the entire community[3]. With an understanding of food-web topology, ecosystem responses to structural change can be predicted. One of the main present-day examples of such change is human-driven biodiversity loss. Suppression of some species from food-webs can lead to a "trophic cascade", where indirect interactions occur throughout the food-web. Studies have also shown food-webs have "rivet-like" thresholds for species removal. This "rivet" model hypothesises that the ecological functions of some species overlap, and so similar to rivets on the wings of a plane, can be removed individually without much change to the overall structure[4]. However, once a threshold is exceeded, there is a dramatic increase in the rate of secondary extinctions when highly connected species are removed[5, 6]. The first objective of this project, in Chapter 3, will be to perform an analysis on the network structure of the Guadalquivir River estuary food-web to understand the ecosystem dynamics. Consisting predominantly of juvenile species, this food-web provides an especially interesting comparison to typical food-webs.

The second objective of the project will be to perform an in-depth analysis of the species abundance time-series data. Abundance measurements are taken at various sites and tides in the estuary. Each of these present different environmental conditions, and we shall examine in Section 4.5 how the species found at each site and tide are correlated. Visualising the data is crucial for understanding the species we are working with and for creating hypotheses. We create a graphical user interface in Section 4.2 to compare species abundances, correlations and monthly average abundances.

The final objective of this project will be to bring this together and use complexity science to determine species interactions using only their abundance. Causality can typically be identified by measuring values such as *Granger Causality*[7]. However, dynamical systems with weak to moderate coupling make this problematic due to their inseparability[8]. Instead, we will apply a technique developed by Sugihara et al.[8], known as convergent cross mapping (CCM). This concept is introduced in Chapter 5 and demonstrated with several examples in Section 5.3 and Section 5.4. Using CCM, we predict food-webs for the dominant species of the estuary, and evaluate them against the actual food-web. With the same model and environmental data for the temperature, salinity and turbidity of the estuary, we shall consider how these environmental factors drive species populations. From the effects of freshwater dam discharges to climate change; understanding interactions between the environment and the ecosystem is fundamental for shaping social and environmental policy.

In Section 5.10 we test for causality between Anguilla anguilla (European eel) and other species. The full life-cycle of Anguilla anguilla is not fully understood, and although it is seen in the Guadalquivir estuary, it is believed to not interact significantly with the ecosystem[2], a hypothesis we shall test.

Chapter 2

Background

2.1 Guadalquivir River Ecological Dataset

The Guadalquivir river estuary is located in the Gulf of Cadiz, in the autonomous community of Andalusia, on the southwestern coast of the Iberian peninsula. The location of the estuary in Western Europe and the waterways in and near the estuary are shown in Figure 1.

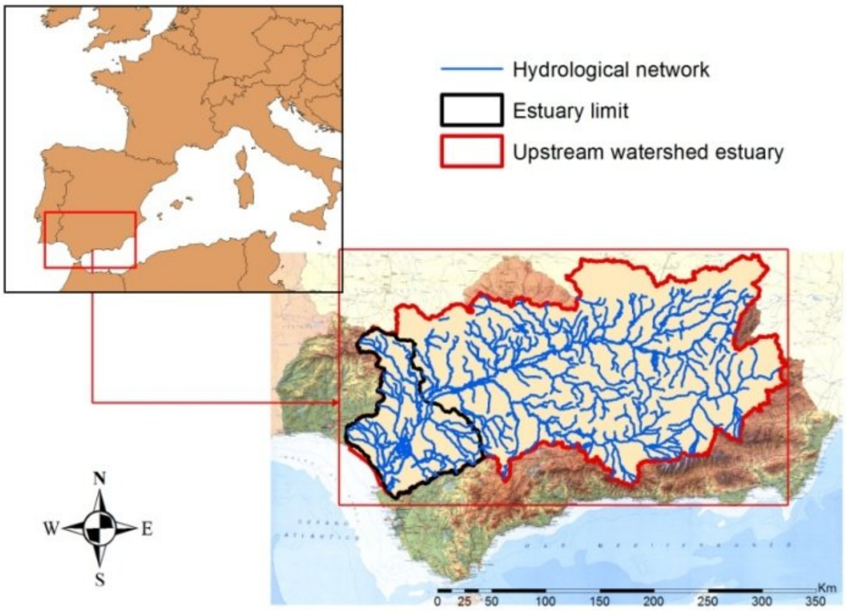


Figure 2.1: Location and map of the Guadalquivir River estuary. Sourced from Carpintero et al.[9]

The fisheries here provide a key natural resource for the area, and influence the local tradition, culture, coastal societies and gastronomy. The combination of these form the main economical activity of the area: tourism. However, this has also led to the area being heavily affected by human activities, including urban concentration, dam regulation, intensive agriculture, contamination and eutrophication, ship transport, and aquaculture[2].

An interesting question to ask is why this area provides such a rich ecosystem of marine life. Sea surface temperature (SST) satellite images have shown the waters in the Gulf of Cadiz to be particularly warm from April to October, with plumes of warm waters exiting the Guadalquivir River[10]. This is due to the land serving as a source of heat during tidal propagation inland[11]. This heating is then further driven by other physical processes such as the greenish colour of the water, water shallowness due to the gulf lying on a wide and shallow platform, increased daylight during spring/summer, and flooding of marshes that have been heated, leading to a greater absorption of heat in the river than offshore. These tidal cycles also lead to the waters being enriched with nutrients[12]. These conditions favour an efficient transfer of primary production towards higher food-chain levels.

Ruiz et al. conducted an investigation into the meteorological and oceanographic factors influencing the population of Engraulis encrasicolus (anchovy) eggs and larvae in the Gulf of Cadiz[13]. This study found the area to be highly favourable for the development of such eggs and larvae due to the warm and chlorophyll-rich waters in the area. However, the study also found that the development of eggs and larvae is negatively impacted at other times of the year by surface currents caused by easterly winds decreasing water temperature. This demonstrates how the Guadalquivir River can provide a favourable home for species, but has many complex factors influencing the cycles of animal populations.

In 1969, Odum explained that an ecosystem can be maintained at an intermediate point in the development of its species by external physical perturbations. This is especially applicable to estuaries and inter-tidal zones, of which the Guadalquivir River has both. In general, these are "maintained in an early, relatively fertile stage by the tides, which provide the energy for rapid nutrient cycling"[14]. With this argument, we can see why the Guadalquivir River estuary might be such an important area for the recruitment of new marine juveniles.

Any model representing the ecosystem will also be made more complex due to the seasonality of the system. With changes in the food-web and species presence over different seasons, we can also expect changes in the interactions between the species. Figure 2.2 shows the observed monthly averages of the temperature, salinity and turbidity. All of these change throughout the year, with decreased rainfall in summer months leading to a lower input of freshwater, causing salinity to increase and turbidity to decrease.

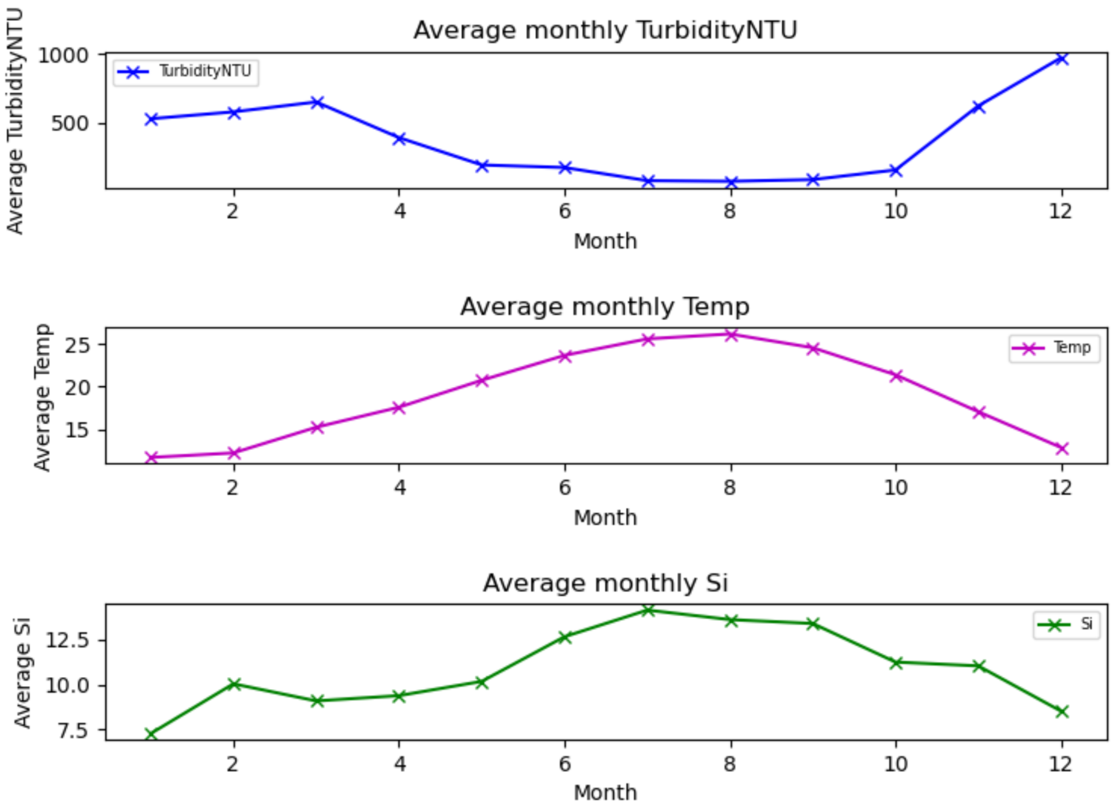


Figure 2.2: Environmental monthly conditions averaged over all sites, tides and years

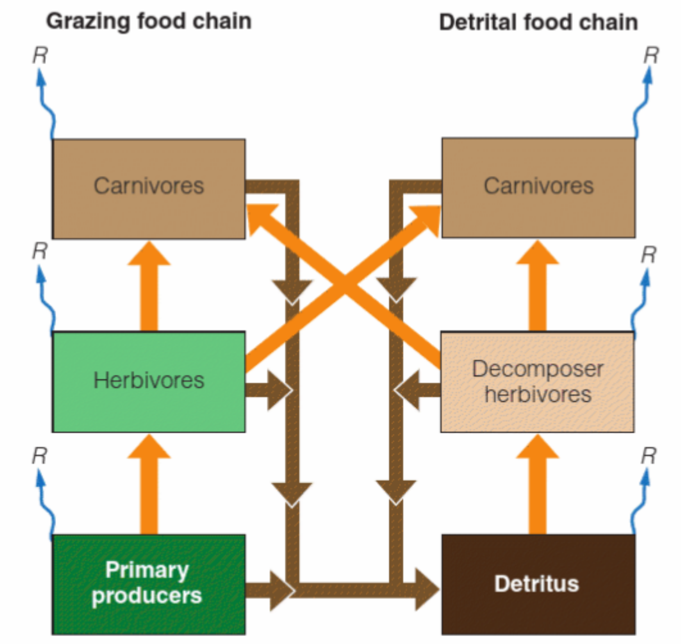


Figure 2.3: A grazing food chain and a detrital food chain. Orange arrows represent the flow of energy due to ingestion, blue arrows represent energy lost due to respiration and brown arrows represent dead organic matter and waste products. Sourced from Elements of Ecology[16].

2.2 Guadalquivir River Food-Web

2.2.1 Food-Web Ecology

Food-webs represent feeding relationships within communities and imply the transfer of food energy from its source in plants, through to herbivores, and then to carnivores. Normally the food-web is constructed from a mesh of food-chains, where a food-chain is a series of arrows representing the flow of food energy[15].

Within any ecosystem, there are two main types of food chains: grazing food chains and detrital food chains, as shown in figure 2.3. These differ in their primary source of energy that the first-level consumers eat. In grazing food chains this source is living plant biomass, while in detrital food chains this is dead organic matter. Zoo-plankton consuming phytoplankton would be an example of a first-level consumer in a grazing food-chain, while bacteria would be a first-level consumer in a detrital food chain[16].

Interestingly, aquatic and terrestrial food-webs have been found to be systematically different in energy flow and biomass partitioning between producers and herbivores, detritus and decomposers, and higher trophic levels[17]. In most terrestrial ecosystems with high standing biomass and relatively low harvest of primary production by herbivores, the detrital food chain is dominant[15]. Meanwhile, in aquatic ecosystems, aquatic herbivores have been shown to accumulate on average three times as much biomass as terrestrial herbivores for a given level of primary production. Conversely, aquatic detritus consumers accumulate a much lower biomass[18]. This implies that the grazing food chain may be dominant here[16].

There are different ways to consider food-webs in terms of what the connections represent. Three possible representations for food-webs, as shown in figure 2.4, are[15]:

1. Connectedness webs which emphasize feeding relationships among species.
2. Energy flow webs to quantify energy flow from one species to another. Thickness of the arrow can represent the strength of the relationship
3. Functional (or interaction) food-webs for representing the importance of each species in

maintaining the integrity of a community and the influence of a species on the growth of others.

For this project, the Guadalquivir River estuary food-web has been constructed as an energy flow web.

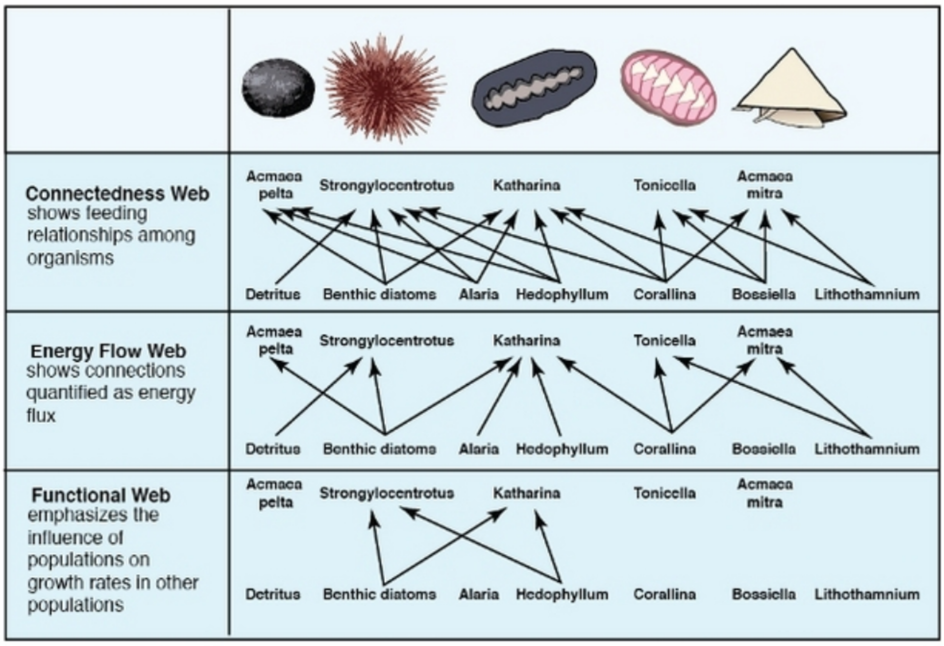

Connectedness Web shows feeding relationships among organisms
Energy Flow Web shows connections quantified as energy flux
Functional Web emphasizes the influence of populations on growth rates in other populations
Acmaea pelta
Strongylocentrotus
Katharina
Tonicella
Acmaea mitra
Detritus
Benthic diatoms
Alaria
Hedophyllum
Corallina
Bossiella
Lithothamnium

Figure 2.4: Three types of food-webs for a rocky intertidal zone on the coast of Washington. *Acemea mitra* can be seen to consume considerable energy but have little influence on the abundance of other species, while *Stronglocentrotus* has significant control on the populations of other species. Sourced from Hui[15].

Food-webs can also illustrate indirect interactions between species. This occurs when one species interacts with a second species but may also influence a third species due to an interaction between the second and third species. A famous experiment by Robert Paine demonstrates this[19]. For two years, he kept an area of shoreline near Washington clear of *Pisaster* (starfish), the top-predator in the area. He observed a decrease in diversity, with the number of species in the system reducing from 15 to 8. This was due to the *Mytilus* (mussles) and *Balanus* (barnacles) dominating the system and excluding some of the other species. The interaction in which a predator enhances less competitive species by reducing the abundance of more competitive species is called keystone predation[16].

Another type of indirect interaction present in some food-webs is apparent competition. This can occur when two species do not compete for limited resources, but still affect each other by sharing the same predator. If the abundance of one species increases, it may increase the abundance of the common predator, which would in turn decrease the population of the other species due to increased predation[16]. This was first demonstrated in an experiment in 1987[20] where sessile bivalves (a type of mollusc) were added to replicate cobble plots. These bivalves share a common set of predators with several mobile gastropods, such as *Panulirus Interruptus*(lobster), *Octopus bimaculatus*(cephalod) and *Kelletia kelletii*(whelk). The addition of bivalves resulted in an increased density of predators and reduced density of gastropods relative to controls. The converse was also true, with higher bivalve mortality in areas where gastropods were more common.

The process of apparent competition relies upon two underlying phenomenona: bottom-up control and top-down control. In bottom-up control, the abundance of any population is limited by the productivity and abundance of the populations in the trophic level below them. The converse of this is top-down control, where predators can control the abundance of their prey[16]. "The world is green" is a famous proposition that introduced the idea of top-down control[21]. Hairston, Smith and Slobodkin argue that "the world is green" because herbivores are controlled by predation, rather than limited food supply. When herbivores are temporarily protected by man, natural events, or are an invasive species, they have been shown to deplete the vegetation in their surrounding area. A famous example of this is the Kaibab deer. In 1906, the area home to the Kaibab deer became protected from hunters when the Grand Canyon National Game Reserve was established. At the same time, many of the predatory mammals of deer were killed by the United States Forestry Service. With fewer predators, the population of the deer greatly increased to 100,000 individuals, which then depleted the herbivory in the range, leading to malnutrition amongst the deer[22].

2.2.2 Food-Web Networks

Food-webs can be represented using directed graphs consisting of S nodes or trophic species. A trophic species is defined to be "functional groups of taxa that share the same predators and prey"[23]. Connecting these nodes are L trophic links or directed edges representing the flow of energy as one species is consumed by another[24].

A characteristic property of some networks is the "small-world" property. The term originates from the "six degrees of separation" social experiment by Stanley Milgram (1967) which shows that most people are six or fewer social connections apart from each other[25]. It is important for understanding and explaining, for example, the rate at which ideas may spread or nodes may influence each other. In network theory, a small-world network is defined to be a network where the expected value for the shortest path between two nodes grows logarithmically as a function of the number of nodes, S. This average shortest path length is called the characteristic path length. The two main properties of small-world networks are then[26]:

- high clustering of nodes compared with a random graph.
- small path length compared to a regular lattice.

The second network characteristic we are interested is whether a network "scale-free". A network is scale-free when the number of edges on each node (the degree of a node) follows a power-law degree distribution. Mathematically, this distribution can be written

$$p_k = Ck^{-\alpha}$$

where p_k is the fraction of nodes that have degree k, and α and C are constants. To detect a distribution like this, one can check if a histogram of the node degrees plotted with logarithmic scales is linear[3].

Some large-scale networks have been reported to self-organize into scale-free networks[27]. Examples of this include the World Wide Web (WWW) or citation patterns in science. Interpreting p_k as the probability that a node in these networks are connected to k other nodes, this probability decays as a power law, following $p(k) \sim k^{-\alpha}$. This has been found to be a consequence of two generic mechanisms in these scale-free networks:

1. networks expand continuously by the addition of new nodes
2. new nodes attach preferentially to sites that are already well connected

Recent theoretical approaches to networks in nature have found them to exhibit small-world and scale-free patterns[28]. For the Guadalquivir River estuary, it will be useful to examine if the food-web follows these characteristics. A recent study by Dunne et al. shows that amongst 17 publicly available food-web networks examined, most do not have small-world or scale-free structure if they exceed a relatively low level of connectance (the fraction of all possible edges that are realised), since they display less clustering than expected[1]. However, the study shows that the degree distribution can be related to the network connectance and size, and that most food-web topologies are still consistent with patterns found in scale-free and small-world networks.

In Chapter 3, we shall examine the Guadalquivir River estuary food-web for these characteristics, as well as analyse its other network properties such as the centrality and clustering.

2.3 Time Series Dataset

A time series is a collection of observations made sequentially through time. In the context of the Guadalquivir River dataset, these observations are of the species biomass and environmental conditions. Denoting the random variable of the observation at time t as X_t, we can create a time series $\{X_t\}$ from these random variables for times $t = 1, ...N$.

The time series is said to be strictly stationary if the joint distribution of $X_{t_1}, ..., X_{t_k}$ is the same as the joint distribution for $X_{t_1+\tau}, ..., X_{t_k+\tau}$ for all $t_1, ..., t_k, \tau$, where τ is called the *lag*[29]. We can then define the autocovariance coefficient at lag τ to be the covariance of X_t with $X_{t+\tau}$

$$\gamma(\tau) = \text{Cov}[X_t, X_{t+\tau}]$$

Standardizing this function, we can produce the autocorrelation function

$$\rho(t) = \frac{\gamma(\tau)}{\gamma(0)}$$

which allows us to measure the correlations between different time points[29].

2.3.1 Detecting Time-Series Causality

Zero Lag Association

A basic way to detect interactions between time-series is by estimating all pairwise Pearson Correlations, with zero lag[7]. For two time-series $\{X_t\}$ and $\{Y_t\}$, this can be done using the formula

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Alternatively, we can use the cross correlation function[30]

$$R_{X_t Y_t}(t_1, t_2) := E[X_{t_1} Y_{t_2}]$$

with $t_1 = t_2$ to have zero lag. An issue with zero lag associations however, is that they can not imply directionality, as this would require a time-asymmetry for information to propagate[7].

Lagged Association

Here, we can now use the lagged cross correlation function by taking $R_{X_t Y_t}(t_1, t_2)$ with $t_1 = t_2 + \tau$ to the cross correlation at lag τ. The value of τ that maximises this function can then indicate the point in time when the time-series are best aligned.

In his thesis, Runge explores the reliability of this method for inferring directionality and physical lags[7]. The thesis analyses a dataset consisting of temperature anomalies in the East Pacific, North Atlantic, Western Europe and Eastern Europe. With a lagged cross correlation analysis, the thesis finds a stronger maximum cross correlation between the East Pacific and North Atlantic (tropical) measurements compared to the European measurements. This suggests a stronger link between the tropical areas, despite them being a much greater distance apart than the European areas. Runge argues that this conclusion is misguided by the influence of autocorrelation between data in the different areas, which is much higher and slower decaying for the tropical time-series.

A different method for measuring the direction of influence between time-series is with Granger Causality. We can say that for time-series $\{X_t\}$ and $\{Y_t\}$, X Granger-Causes Y if

$$\sigma^2(Y_t | U_t^-) < \sigma^2(Y_t | U_t^- \setminus X_t^-)$$

where $\sigma^2(Y_t | U_t^-)$ is the variance of the residual predicting Y using the information in the entire universe until the present, U_t^-. $\{U_t^- \setminus X_t^-\}$ is the set $_t^-$ excluding the points in set X_t^-. When this statement is true, it implies there is some unique information in X about Y[7]. This relies on the assumption that information about $\{X_t\}$ can be formally removed from $\{Y_t\}$.

Convergent Cross Mapping

Unfortunately, Granger Causality based methods have been shown to be inconsistent when applied to ecosystems[8]. This is due to the assumption that information about X can be separated from Y. In a dynamic system like an ecosystem, where variables have weak to intermediate coupling, the entire system must be considered as a whole. As such, we use the convergent cross mapping (CCM) method introduced by Sugihara et al.[8]. Full details of this method are explained in Chapter 5.

2.4 Data Collection

Both the food-web and species abundance datasets used in this report have been provided by Cesar Vilas[2]. The food-web was constructed from multiple studies of species interactions, typically finding interactions by direct observation of fish guts to determine what species have been consumed by the observed species. The species abundance time-series data is from a long-term observation of the Guadalquivir River estuary under the LTER program. Starting in June 1997, and with data available to us until November 2018, 5 sites were sampled at 4 different tides each month. There is a gap of 34 months in the data collection, between September 2015 and July 2018, so to avoid issues that this may cause, we choose to truncate the data we use at September 2015. For this project, we will label the sites

$$\{1, 2, 3, 4, 5\}$$

in order of distance from the sea, with site 1 being furthest downriver and site 5 being furthest upriver. The tides will be labelled

$$\{1, 2, 3, 4\}$$

in chronological order with

- tides $\{1, 3\}$ being flood tides
- tides $\{2, 4\}$ being ebb tides

For each sample, fish abundances and environmental conditions were recorded. After the first 24 months, only sites 3 and 5 were sampled, since they seemed to be representative of the other sites nearby them. In Chapter 4, we test this hypothesis by examining the correlation of species abundances between sites.

Since for the purposes of this report we do not have full knowledge of the techniques used to record the data, we therefore do not know the error margins of the data. We shall bear this in mind but will not consider measurement accuracy where it concerns the raw data in this report.

Chapter 3

Food-Web Analysis

3.1 Objectives

We begin with a network analysis of the food-web. The objectives here are to identify the characteristics of the Guadalquivir river estuary food-web. The structure of ecological food-webs is still a point of discussion, so we will be looking to compare the Guadalquivir estuary food-web characteristics to the current accepted theories on food-webs. In particular, we will consider if the food-web exhibits small-world and scale-free structure, which Dunne et al. argue do not happen if a relatively low level of connectance is exceeded[1]. By analysing network centrality, we will also be able to identify the prey and predators core to the ecosystem, and with the degree of the nodes we can attempt to identify the keystone species.

3.2 Initialisation

To prepare the data, we import the food-web CSV data, provided by Cesar Vilas[2], as an array into Python using the "csv" package. The food-web is represented by an adjacency matrix, where each food-web node is represented by a row and a column. It consists of binary elements, either "0" or "1", where a "1" represents the species in the row having been found to prey on the species in the column, and a 0 represents no such interaction having been found. If the species are labelled as integers from 1 to n, we can define the directed food-web adjacency matrix as

$$D_{ij} = \begin{cases} 1 & \text{if there is an edge from species } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

Initially, we make this matrix symmetric so that a "1" represents any sort of predator-prey interaction and the food-web network can be considered undirected. This is done to increase the simplicity of the initial analysis, since when measuring network properties an undirected network has less edge cases - some of which are detailed in Section 1.9. The initial food-web prediction is done for an undirected food-web since it is easier to predict an interaction without having to specify the direction of the interaction. The analysis on food-web characteristics by Dunne et al. is also done on undirected food-webs, so an undirected analysis allows for a direct comparison[1]. We symmetrize the adjacency matrix by taking the maximum of each entry and the corresponding entry in a transpose of the matrix. Taking D as the directed matrix and U as the undirected matrix, this can be expressed as

$$U_{ij} = \max(D_{ij}, D_{ji})$$

3.3 Visualisation

The popular Python libraries networkx and matplotlib allow us to create a visualisation of the food-web, shown for the undirected network in figure 3.1.

We can see that there is a large amount of connectance between the large amount of nodes in the centre of the figure, but that there are also some nodes with only one or two edges.

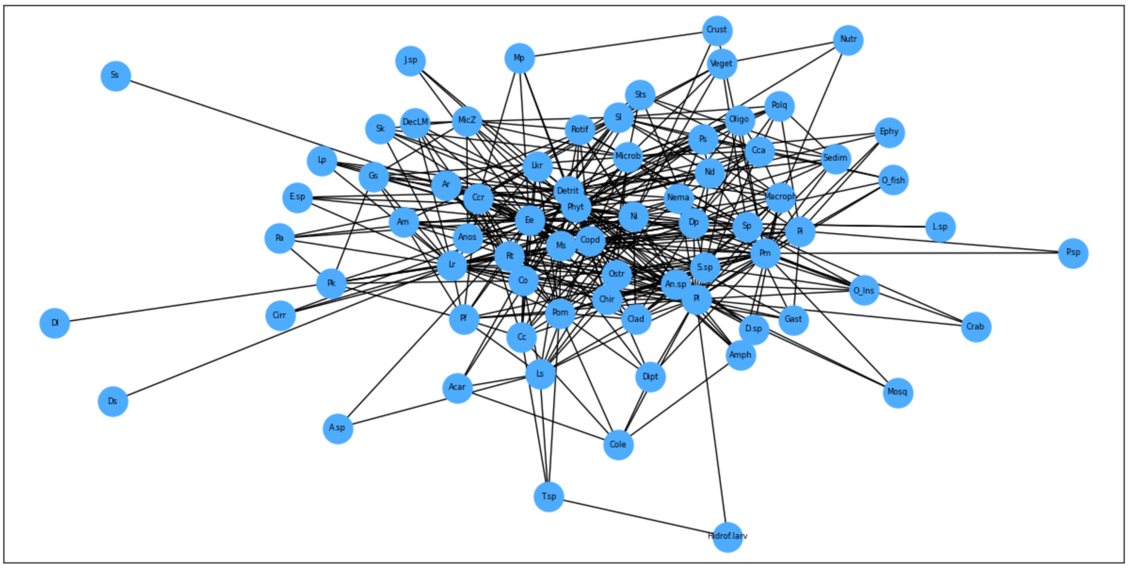


Figure 3.1: Undirected food-web visualization using networkx. Nodes are labelled using short names provided in the species list[2].

Throughout this food-web analysis, networkx library functions were also used to verify calculated values such as the average degree, clustering coefficient and centrality, in order to improve reliability of results.

3.4 Undirected Food-Web Degree

The degree of a node for an undirected graph is defined as the number of edges connected to it[3]. Since we are working with a simple graph and two nodes can only be connected at most by one edge, the degree in this case is also the number of neighbours a node has. We measure this by summing the total interactions for each species' row in the undirected food-web matrix. Mathematically, this can be expressed as

$$k_i = \sum_{j=1}^{N} U_{ij}$$

where k_i is the degree of node i, and there are N nodes in the network.

With this, we can calculate the average degree, c of all nodes, defined as

$$c = \frac{1}{N}\sum_{i=1}^{N} k_i$$

Since each edge increments the degree of two nodes, we can also consider the average degree in terms of the number of edges, M as

$$c = \frac{2M}{N}$$

When calculating degree with the undirected network matrix, we must make sure to also count loops (edges from a node connecting to itself) in a consistent way, with a single loop causing an increase of 2 in the total degree. We allow loops in the network we construct since some species, such as *Engraulis encrasicolus* (European anchovy), have been recorded to consume fish of their own species. Since taking the transpose of the directed food-web matrix will not count loops (1s on the diagonal of the matrix) twice, we correct for this when calculating the degree.

In food-web topology, networks are often considered in terms of links per species and connectance. With N species and M edges, the links per species is $\frac{M}{N}$, which will be smaller than the average undirected degree by a factor of 2. Connectance measures how many of all possible links are realized with the formula $\frac{M}{N^2}$[1].

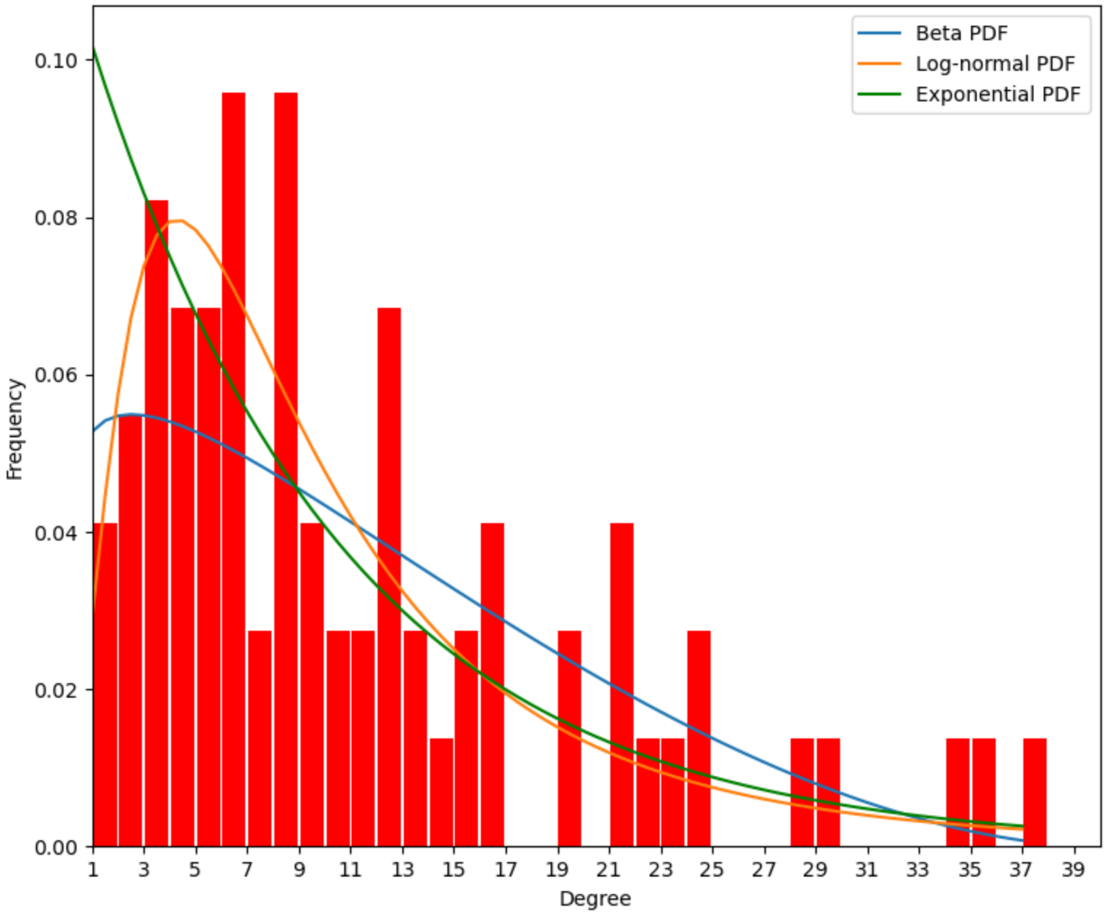


Figure 3.2: Undirected food-web degree histogram

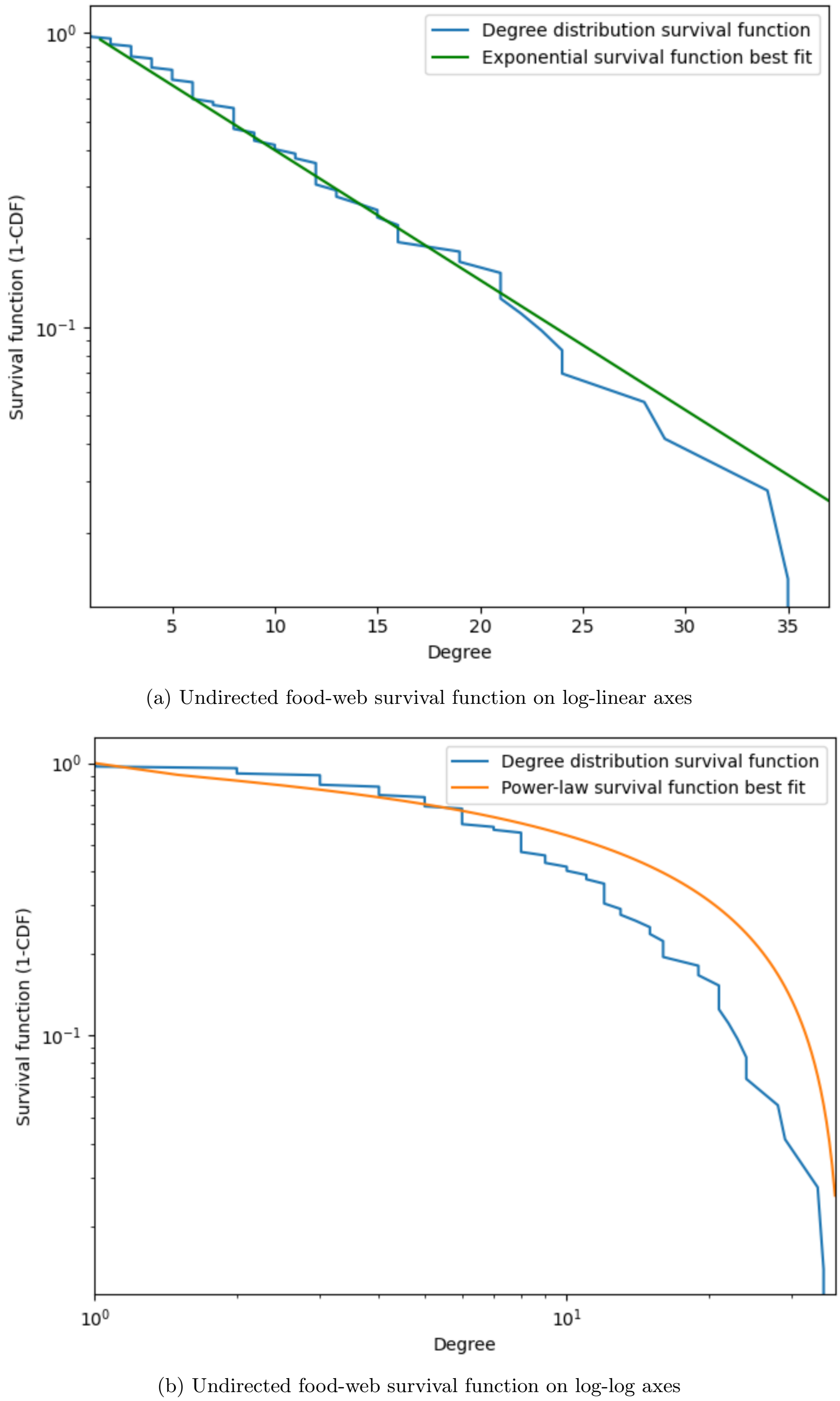
The histogram in figure 3.2 shows the frequency of each degree. We can see that a large number of species have just a few interactions, with the number of interactions decaying at a rate that seems exponential. Attempting to fit beta, log-normal and exponential probability density functions using the scipy Python library, we find that the exponential distribution seems to perform best, but none of them match the shape exactly. This is due to some of the histogram bins being empty and having zero probability mass, most likely due to the small distribution sample size.

In order to avoid this problem, as in Dunne et al.[1] and Montoya et al.[28], we use the cumulative degree distribution to give a more accurate illustration of the distribution shape. We plot the survival function, $1 - CDF(k)$, of the undirected degree distribution on log-linear axes. The shape of the function determines the distribution by being

- a uniform distribution if upwards curving
- an exponential distribution if straight
- a power-law distribution if downwards curving

We find that it gives a straight line and when we plot the survival function of the exponential distribution that was fitted in figure 3.2, it follows the shape of the food-web survival function well, as shown in figure 3.3a. If the degree distribution survival function gives a straight line on a log-log plot, this will suggest that it follows a power-law distribution. This graph is shown in figure 3.3b, with a plot of the survival function of the power-law distribution that was fitted in figure

3.2. We find that the degree distribution survival function does not produce a straight line, and does not follow the power-law survival function well. From this, we conclude that the degrees of the undirected food-web are exponentially distributed, and hence is not scale-free since it doesn't follow a power-law distribution.



(a) Undirected food-web survival function on log-linear axes

(b) Undirected food-web survival function on log-log axes

Figure 3.3: Undirected food-web survival function

The degree of a node can be used as a measure of centrality - the "importance" of a node in a network. The species with the highest degree centrality is *Engraulis encrasicolus* (European anchovy), with a degree of **37**. This suggests that the European anchovy is a keystone predator for this ecosystem, interacting with 50.6% of species.

The minimum degree is shared amongst several species, that have a degree of **1**:

- *Diplodus sargus* - Sargo (fish)
- *Dicentrarchus labrax* - European bass (fish)
- Other insects

Having only one interaction suggests that all of the above should be either top predators or primary preys.

The links per species in the river estuary is **5.452** and connectance is **0.0747**. This ranks in the middle of the 16 food-webs studied in Dunne et al., which had connectance values ranging from 0.026 to 0.0315[1]. The 16 food-webs in Dunne et al. showed a higher mean connectance of 0.11. However, the standard deviation of these 16 food-webs was very high, at 0.09, so the Guadalquivir estuary food-web lies within one standard deviation from the mean of the 16 food-webs. The lower connectance could potentially be explained by the fact that the Guadalquivir estuary is home to predominantly juveniles which may interact less than adults, or that there are species interactions that have not been noticed and recorded yet in the food-web.

3.5 Closeness

Closeness is a measure of the distance from a node to all other nodes. We measure it with the shortest path length, d_{ij}, which is the minimum number of edges that must be followed to reach node j from node i. To calculate the shortest paths between all nodes, we run the Floyd-Warshall algorithm[31] to obtain them in $O(n^3)$ time complexity. In summary, the algorithm recursively finds the shortest distance, l_{ij}, between nodes i and j by considering the shortest path that passes through intermediary node k. After testing all possible intermediary nodes, the intermediary node that gives the shortest distance is used. If $l_{ij}^{(m)}$ is the shortest path from i to j in at most m steps, the recursive case can be expressed as[32]

$$l_{ij}^{(m)} = \min\{l_{ij}^{(m-1)} + w_{kj}\}$$

where w_{kj} is an element of the adjacency matrix, which in our case will be constructed to be 1 if an edge exists, and infinity otherwise. The base case for the recurrence is then

$$l_{ij}(0) = \begin{cases} 0, i = j \\ \infty, i \neq j \end{cases}$$

The Python code implementation for the algorithm can be found in the source code archive.

Taking the mean of all pairs of shortest path lengths, we get an average shortest path length of **2.078**. Comparing this to the average shortest path lengths of ecosystems in Dunne (2002) which had values between 1.33 and 3.74, we find that the Guadalquivir estuary has a relatively low average shortest path length. This means that most nodes are closely connected and is consistent with small-world topology.

Closeness can also be used as a measure of centrality. In order to have a higher score for nodes that are a short distance away from other nodes, the closeness centrality score, Cl_i, takes the inverse of the mean closeness to all nodes

$$Cl_i = \frac{N}{\sum_{j=1}^{N} l_{ij}}[3]$$

The 10 most central nodes in terms of closeness are shown in table 3.1. We can see that all of these nodes have a high degree relative to the network mean of **10.82**, and can suggest that these species are likely to be keystone species. In Chapter 3, we predict the most dominant species in the ecosystem using the abundance dataset, with the results shown in table 4.1. A list of the dominant species from the perspective of an ecologist[2], is also shown in table 4.2. We find that the species in table 3.1 with high closeness centrality are very similar to the other predictions of the dominant species.

Node	Clustering Coefficient	Degree
Palaemon longirostris	0.658	35
Engraulis encrasicolus	0.652	37
Liza ramada	0.619	29
Phytoplankton	0.608	34
Neomysis integer	0.603	23
Copepoda	0.593	21
Mesopodopsis slabberi	0.593	21
Crangon crangon	0.579	24
Pomatoschistus species	0.575	21
Dicentrarchus punctatus	0.570	22

Table 3.1: Largest 10 closeness centrality scores

3.6 Clustering Coefficient

The clustering coefficient of a network measures the extent to which nodes in the network tend to cluster together, by examining the amount nodes with a common neighbour are directly connected themselves. It can be calculated at a global level by considering how many paths of length two are closed. If there exists a path between a triple of nodes (u, v, w), with edges $(u, v), (v, w)$, then we can see that these nodes are connected. If there also exists an edge (u, w) then these nodes form a triangle and we can say that these nodes are closed, as shown in figure 3.4. The global clustering coefficient can therefore be defined as

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets}}[3]$$

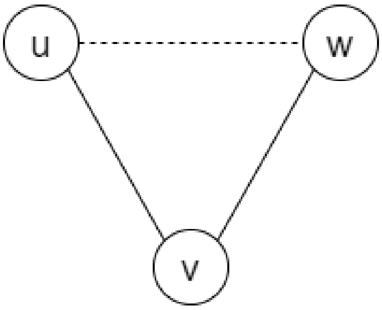


Figure 3.4: Connected triplet of nodes (u, v, w) that becomes a closed triangle when the dashed edge is realized.

Alternatively, the clustering coefficient of a network can be measured by averaging the local clustering coefficients of each node. This is the approach that we shall choose to take as it will also provide information about which species exhibit stronger clustering. For a node i, this can be defined as

$$\begin{aligned} C_i &= \frac{\text{number of connected pairs of neighbours of } i}{\text{number of pairs of neighbours of } i} \\ &= \frac{2 \times \text{ number of connected pairs of neighbours of } i}{k_i(k_i - 1)}[3] \end{aligned}$$

where k_i is the degree of node i. The second equation follows the fact that node i has $\frac{1}{2}k_i(k_i - 1)$ pairs of neighbours. We also define $C_i = 0$ for $k_i = 0, 1$, to ensure the equation is well defined.

The network average clustering coefficient is then[26]

$$\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i$$

It must be noted that this definition is not equivalent to the global clustering coefficient. Using this formula, the network average clustering coefficient that we calculate for the Guadalquivir estuary is **0.2556**. This is significantly higher than most of the clustering coefficients of ecosystems studied in Dunne et al.[1], and is consistent with the characteristics of a small-world network. However, it is difficult to directly compare the clustering coefficients of networks since the clustering coefficient is not constant with respect to network size. We will instead later compare this clustering coefficient to random networks of the same size.

The node with the highest clustering coefficient was the phylum *Rotifera*, with a coefficient of **0.6071**. This implies that the 8 species that consume or are consumed by Rotifers also have a high probability of interacting with each other. The species with the second highest clustering coefficient was *Argyrosomus regius* with a clustering coefficient of **0.600** and degree of 10. *Argyrosomus regius* is also a top predator, with no other species preying on it, so we can conclude that its preys are likely to interact with themselves.

Node	**Clustering Coefficient**	**Degree**
Rotifera	0.607	8
Argyrosomus regius	0.600	10
Aphia minuta	0.528	9
Sigara species	0.485	12
Decapoda	0.467	6
Paragnathia formica	0.455	11
Cladocera	0.449	13
Cyathura carinata	0.436	11
MicZ	0.429	8
Cyprinus carpio	0.409	12

Table 3.2: Largest 10 clustering coefficient scores

3.7 Comparison to Random Networks

In a similar manner to Dunne et al., we compare the estuary food-web to random networks[1]. A network can be said to be scale-free if it has:

- a higher clustering coefficient compared to a random network
- shorter average path length compared to a regular lattice

As such, we will generate uniformly distributed random networks to compare the Guadalquivir estuary's properties to, in order to determine if it is small-world.

Initially, we construct a random network to have the same number of nodes and edges as the Guadalquivir estuary, with the edges uniformly distributed. We use equal numbers of nodes and edges as clustering and shortest path length are both distributed as functions of the degree and number of nodes. We generate the random food-web by taking the following steps:

1. Generate 73 nodes represented by a 73×73 adjacency matrix.
2. Uniformly distribute an edge by randomly selecting two nodes to connect. If the two nodes are already connected, keep randomly selecting pairs of nodes, until an unconnected pair is found.
3. Repeat step 2 until there are L edges.
4. Perform a breadth-first search to verify the graph is weakly-connected. A graph is weakly-connected when any node can be reached by following undirected edges.

If the network is not weakly-connected, it is rejected and the network construction is repeated. As done in Dunne et al[1], we want to generate weakly-connected networks in order to increase similarity to our data which lacks disconnected species. The properties of the random network were then measured as before, and averaged over 100 iterations.

From 100 simulations of a random food-web, we have the same number of links per species as expected by construction. The average clustering coefficient of the random networks is **0.151**, while the Guadalquivir estuary food-web's clustering coefficient is 0.256. This shows that the Guadalquivir estuary food-web has significantly higher clustering than a uniformly distributed random network, a property required for the food-web to be small-world.

The average shortest path length on the random food-webs is 2.017, slightly larger than the Guadalquivir estuary food-web's characteristic path length of 2.078. However, by the definition of small-world requires the comparison of shortest path length to a regular lattice rather than a random network.

Therefore, we also construct a regular lattice. The regular lattice we create is a circular network, where each node is connected to d of the next nodes in the circle. Since every node must have the same number of edges, we use 5 links per species to make the regular lattice as similar to the Guadalquivir estuary food-web as possible. This gives an average degree of 10 (since each edge is counted twice) and 365 total edges between 73 nodes. Since this construction is deterministic, we do not do multiple iterations. The average shortest path length of this regular lattice is **7.708**, a value significantly higher than that of the Guadalquivir estuary food-web. Therefore, the Guadalquivir estuary food-web also shows small-world characteristics with respect to shortest path length.

3.8 Comparison to Scale-Free Networks

Although we have seen that the undirected degree of the network is not scale-free and follows an exponential distribution, rather than a power-law distribution, we decided to investigate how the food-web compares with networks that are scale-free.

We construct these networks using the two mechanisms found to create scale-free degree distributions: growth and preferential attachment. Growth is where the number of nodes continuously increases, and preferential attachment is where new nodes are more likely to be connected to nodes with higher degree. A model that generates such networks that we choose to use is the Barabasi-Albert model (BA model)[27]. The algorithm for such a model, where each node adds m edges, is:

- Start with an arbitrary configuration of m_0 nodes.
- Generate a new node and create m edges.
- Connect each edge to a node i with probability $\frac{k_i}{\sum k_j}$ where k_i is the degree of node i and the sum is over all existing nodes.

Since each node adds multiple edges simultaneously, we also check that none of these edges go to the same edge since we do not allow multi-links, to better resemble a food-web. If this check fails, then the edges are generated again. This leads to some looping initially when the few nodes result in a higher chance of edges choosing the same node, but does not seem to impact performance substantially over a time-scale of 100 iterations.

We run 100 iterations of generating a scale-free network using the Barabasi-Albert model, with an initial configuration of 5 nodes, each connected to the first node and the previous node. This initial configuration is shown in figure 3.5. For each new node, as with the random networks, we create 5 edges that are now attached using preferential attachment. Creating a network with 73 nodes gives an average undirected degree of **9.507**. The average shortest path length is **2.098** and the average clustering coefficient is **0.234**. Compared to the estuary food-web, the average shortest path length is very similar, and the average clustering coefficient is slightly lower. It is important to note in this comparison that the average degree is not exactly the same as the estuary food-web and the starting configuration is arbitrarily defined. Both of these will have an effect on the average shortest path length and clustering coefficient. Nevertheless, we can conclude that the estuary food-web shows some characteristics that are similar to randomly generated scale-free networks. Scale-free networks have been shown to be robust to a failure of a fraction of nodes, and so we hypothesis that the Guadalquivir estuary may also show similar robustness.

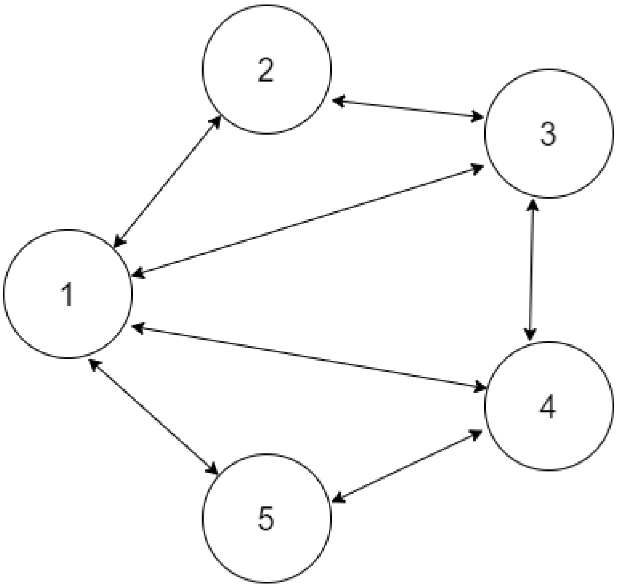


Figure 3.5: Initial configuration for Barabasi-Albert model

3.9 Directed Food-Web

The food-web can also be considered as a directed network. We decide to adopt the convention whereby a directed edge represents the flow of energy, and so an edge from X to Y represents Y consuming X. The degrees of the directed food-web are shown plotted as histograms and survival functions in figure 3.6. The out degree of the node is the number of outgoing edges from that node, while the in degree of the node is the number of incoming edges. We see that the degree distribution in both cases is similar, with the majority of nodes having a small degree, but several nodes with a much larger degree. When considering the cumulative distributions, we see that the distributions are best fitted by an exponential distribution in both cases. This suggests that neither the degree-in nor the degree-out are scale-free.

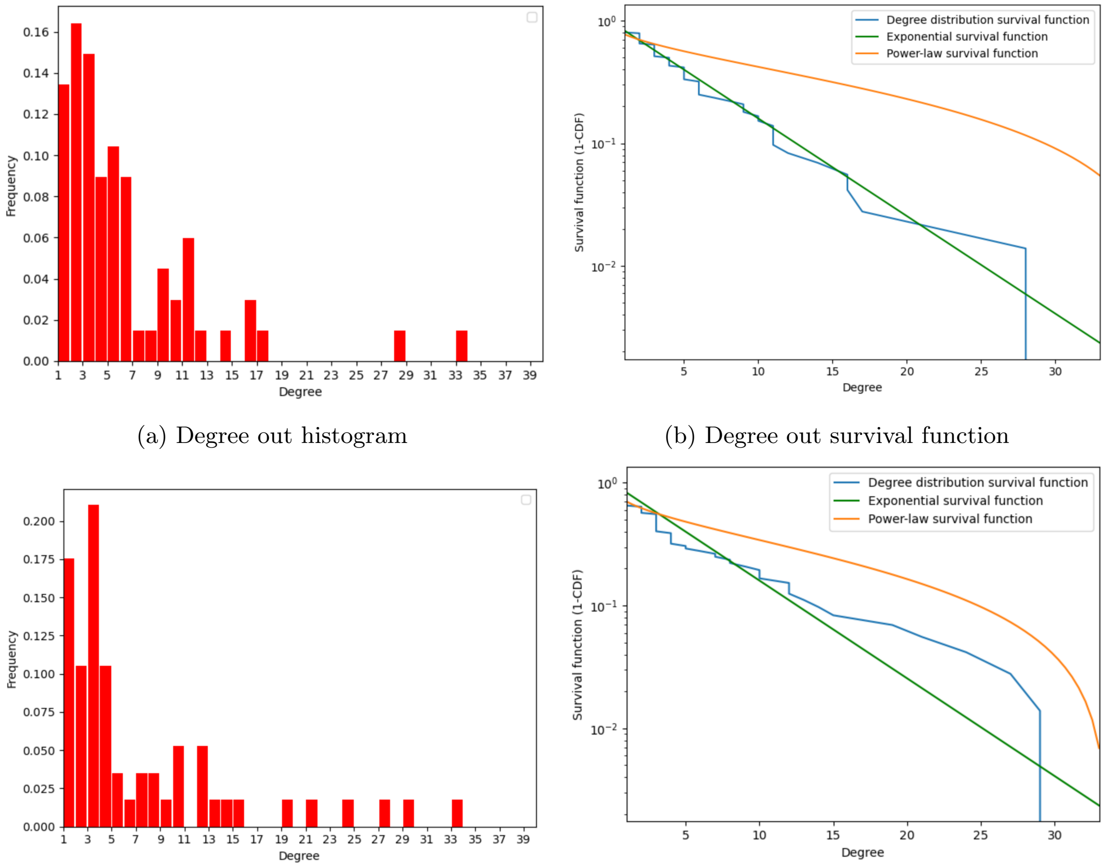
The species preyed on by the most species, with the highest degree out, is *Phytoplankton* with a degree of 33. The species that preys on the most species, with the highest degree in, is *Engraulis encrasicolus* with a degree of 33. The average degree is **5.452**. This is the same for both the degree in and out, since it is a function of the number of edges and nodes.

The average shortest path length in the directed food-web is **0.345**. This is much lower than the average shortest path length in the undirected case. However, this is due to the convention that if no path exists between two nodes, the distance is set to 0, and so with fewer connected nodes, the average shortest path length is reduced. A regular lattice made of 73 nodes, where each node has a directed (ingoing or outgoing) degree of 5, has an average shortest path length of **4.11**. This is much larger than the directed food-web, but this comparison is unreliable to make since the regular lattice is fully connected, whereas the food-web is not.

When a graph is disconnected, the normal convention is to change the formula to use the sum of the reciprocal of the shortest path length to other nodes, and replacing $\frac{1}{0}$ in the sum to be 0, ignoring the disconnected nodes[33]:

$$Cl_i = \sum_{\substack{j=1 \\ j \not\subset D_i}}^{N} \frac{1}{l_{ij}}$$

where D_i is the set of node indexes not connected to i. Using this formula, we find the node with the greatest closeness is *Phytoplankton* with a closeness of **42**, when shortest path lengths are considered with edges leaving prey. When edges are considered entering prey, the node with the greatest closeness is *Liza ramada* with a closeness of **47.17**. Thus, we can conclude the prey that is closest to all other species is *Phytoplankton*, and the predator closest to all other species is *Liza ramada*.



(a) Degree out histogram

(b) Degree out survival function

(c) Degree in histogram

(d) Degree in survival function

Figure 3.6: Directed food-web degree histograms and survival functions

When considering the clustering coefficient for a directed network, the neighbourhood of a node is any other node with an edge coming in or out of the node. Since the network is now directed, a pair of neighbours can now be potentially connected two ways: once in each direction. Therefore, a node now has $d_i^{tot}(d_i^{tot} - 1)$ potential pairs of neighbours, whhere d_i^{tot} is the sum of the degree in and degree out of node i. However, to not count "false" triangles where a pair of directed edges connect to the same node, we subtract these edges from the denominator[34]. Let $d_i^{\leftrightarrow}$ be the number of nodes connected to node i in both directions. The directed clustering coefficient is then

$$C_i = \frac{\text{number of connected pairs of neighbours of i}}{d_i^{tot}(d_i^{tot} - 1) - 2d_i^{\leftrightarrow}}$$

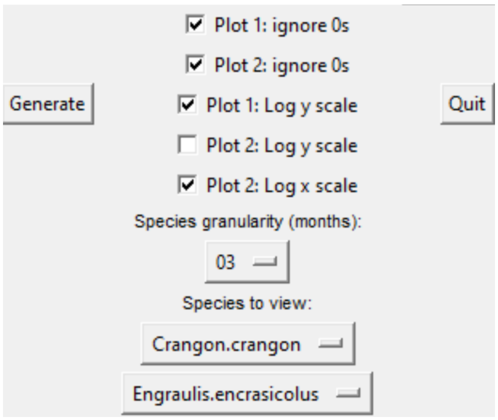
The average clustering coefficient for the food-web is **0.135**. 100 randomly generated directed networks with a uniform degree distribution give an average clustering coefficient of **0.0727**. Therefore, with a larger clustering coefficient compared to a random network and a shorter average path length compared to a regular lattice, we find that the estuary food-web is also small-world like when considered as a directed network.

Chapter 4

Species Abundance

4.1 Introduction

In order to robustly utilise and process the abundance data, we first had to explore and understand the dataset. We started by creating a simple graphical user interface, shown in figure 4.2, to quickly view and compare the population time-series of two species. Noticing many species with few non-zero abundance measurements, we analysed which species we should focus on. Finally, we investigate the measurements taken at different sites and tides to determine if we can average or sum the data between these sites and tides.



☑ Plot 1: ignore 0s

☑ Plot 2: ignore 0s

☑ Plot 1: Log y scale

□ Plot 2: Log y scale

☑ Plot 2: Log x scale

Species granularity (months): 03

Species to view: Crangon.crangon

Engraulis.encrasicolus

Figure 4.1: Graphical user interface for visualising species time-series

4.2 GUI

The graphical user interface allows selection of two species from a drop-down menu and some customisation options for the plots. The generated plots for two dominant species, *Crangon crangon* (common shrimp) and *Engraulis encrasicolus* (European anchovy), is shown in figure 4.2. Plotting the raw abundance data of two species on a plot produced a very busy time series that was difficult to read. Due to the large difference in abundance between different seasons for most species, logarithmic axis are required for displaying the abundance. Some species also have many measurements where they are not present and have an abundance of zero. This creates a lot of noise visually, and so for clearer plots, these zero abundance measurements can be ignored in the GUI configuration options.

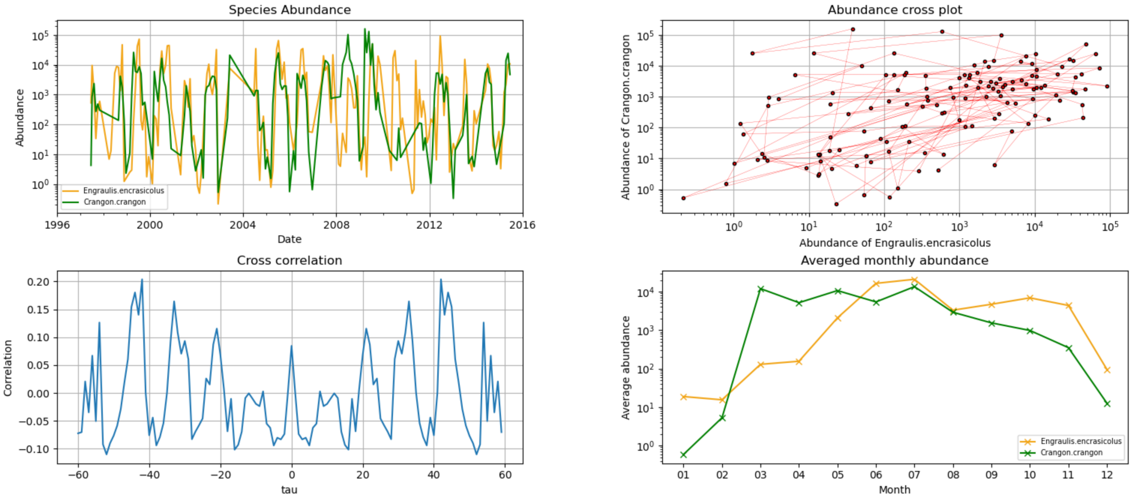


Figure 4.2: Abundance plots of *Crangon crangon* (common shrimp) and *Engraulis encrasicolus* (European anchovy). Log axes are used for representing abundance and measurements of zero are ignored

Plotting the abundance of both species as simple time-series in the first plot allows for a fast visual interpretation of how the abundances of the species may compare, and how the abundances varied over the entire observation period.

In the second plot, plotting both abundances as a plot against each other allows testing to see if there may be visible relationship or pattern between the two species abundances. For example, some species may have a linear relationship where as one abundance increases, so does the second, producing a straight line on the cross-plot. This lines on the plot trace the phase space of the system of these two species.

The third plot illustrates the cross covariance function between the two abundances. The values of the first time series are shifted by a value τ (representing months), and the Pearson correlation is then calculated between the two series, using the formula

$$\rho = \frac{\langle ab \rangle - \langle a \rangle \langle b \rangle}{\sigma(a)\sigma(b)}$$

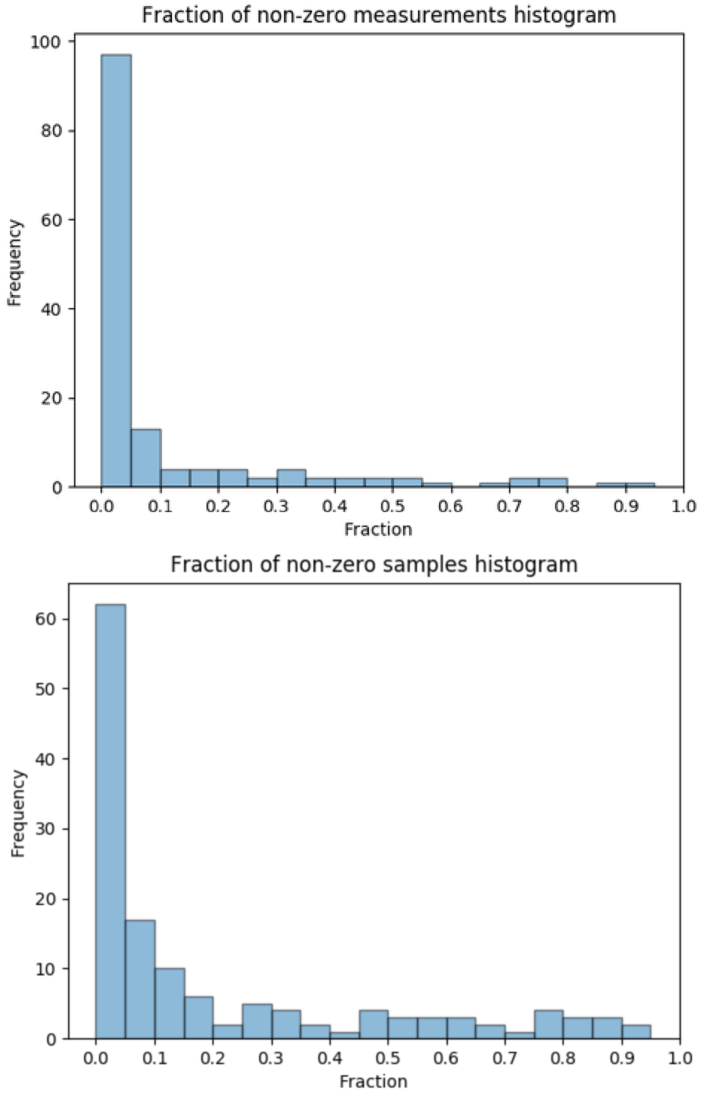
where a and b are the two time series, $\langle . \rangle$ represents the mean, and $\sigma(.)$ represents the standard deviation. The cross correlation plot shown in figure 4.2 shows that the two species seem alternate being somewhat correlated and uncorrelated every 6 months.

The final plot shows the abundance of the species at each month, averaged over all observed years. This allows for simple checking of the seasonality of the abundances, with 4.2 showing that *Crangon crangon* is most abundant in spring, and *Engraulis encrasicolus* is most abundant in late summer.

4.3 Measurements with Zero Abundance

The exploratory abundance plots showed that many species have very low, and often zero abundance for most measurements. For modelling interactions between species, we prefer species that are consistently observed in the ecosystem, rather than just having a seasonal presence, as this provides more usable data points. Therefore, rather than viewing species with high total abundance as "high quality" species, we prefer species with less non-zero abundance measurements.

The large amount of zero abundance measurements is further demonstrated by calculating the fraction of non-zero measurements for each species and plotting a histogram, as shown in figure 4.3. Here we consider a measurement as an individual observation of the abundance, at a specific tide and site, while a sample consists of all measurements done for a species in a particular month, across all sites and tides observed. The histograms highlight that the majority of species are present in less than 10% of observations, and that in order to reduce the number of zero abundance measurements, it's much better to consider abundance in terms of samples. When considering non-zero samples, an upper quartile of 0.448 is achieved, along with a lower quartile of 0.0149. If we were to consider just the top 25 species, in terms of non-zero samples, this mean all of them have at least 63.5% non-zero samples.



Fraction of non-zero measurements histogram

Fraction of non-zero samples histogram

Figure 4.3: Histograms of the fraction of non-zero measurements and samples for all species

According to Cesar Vilas[2], the estuary is predominantly occupied by several majority species. This is in agreement with the data presented in figure 4.3. With respect to fraction of non-zero abundance samples, we can calculate from the data the top 10 species that we expect to have a constant presence in the estuary. The table of these species is shown below in table 4.1. *Anguilla anguilla* is present in this top 10 but since it is not included in the food-web, we shall ignore it for now. However, a further investigation is done into *Anguilla anguilla* in a later section.

Species	**Fraction non-zero samples**
Pomatoschistus	1
Palaemon longirostris	1
Pomatoschistus minutus	0.9908
Mesopodopsis slabberi	0.9862
Neomysis integer	0.9862
Engraulis encrasicolus	0.9771
Crangon crangon	0.9725
Dicentrarchus punctatus	0.9541
Rhopalophthalmus tartessicus	0.9450
Liza saliens	0.9174

Table 4.1: Species with the 10 largest fractions of non-zero samples

Mesopodopsis	Palaemon macrodactylus
Rhopalophthalmus	Crangon crangon
Neomysis	Dicentrarchis punctatus
Engraulis encrasicolus	Argyrosomus regius
Pomatochistus	Liza species
Sardina pilchardus	Chelon labrosus
Palaemon longirostris	Anguilla anguilla

Table 4.2: Most abundant estuary species as reported by Cesar Vilas[2]

We can see that all the species in table 4.1 fall into a subset of the species in table 4.2, reported by Cesar Vilas[2] to be abundant in the estuary.

4.4 Seasonality of Species Abundances

Another scope with which we viewed the species abundance data was in terms of the seasonality of the various species. We created a grid of plots of the monthly average species abundances, as shown in figure 4.4. This was created by calculating for each species the total abundance for each month, averaged over all observed years. The purpose of this was to convey and easily compare the seasonal patterns of each species. We can see that while a majority of species experience a peak in abundance over the summer, this is not true for many species, with several having a higher abundance in winter months.

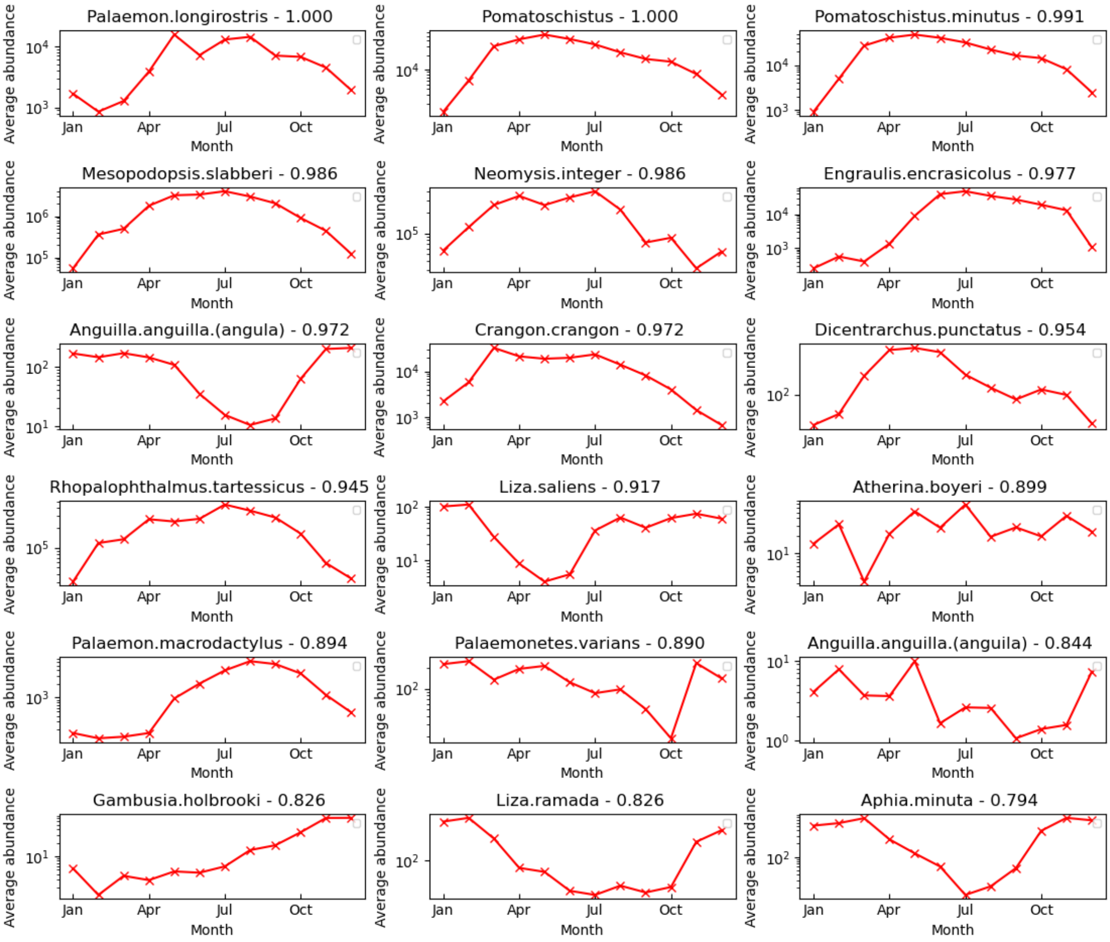


Figure 4.4: Monthly average abundance for the top 18 species. Title of each plot contains fraction of non-zero samples for that species.

4.5 Site & Tide Correlation Heatmap

Data for all 5 sites only exists for the first 24 months of observation. After this period, observations were reduced to only sites 3 and 5 as these seemed to represent the other sites well, with site 3 similar to site 1 and 2, and site 5 similar to site 4[2]. A simplified representation of these site locations is shown in figure 4.5. We investigated the correlation between the sites to verify this hypothesis, and also to explore if it is possible to combine the data from different sites, for example by averaging abundances between sites 1, 2 and 3 where data for all sites exists, to improve data quality by using more samples.

The correlation between two sets of measurements at two sites and tides was calculated by finding the Pearson correlation for each species at the two sites and tides, and taking the mean. As discussed before, a large amount of species had many measurements of 0 abundance. Therefore, we restricted the correlation measurement to the top 25 species, with respect to the fraction of

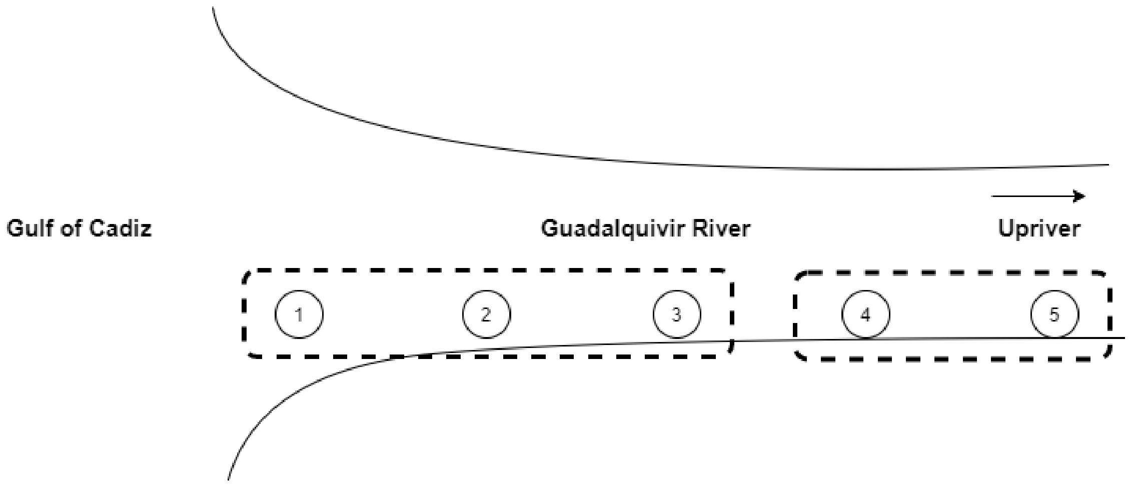
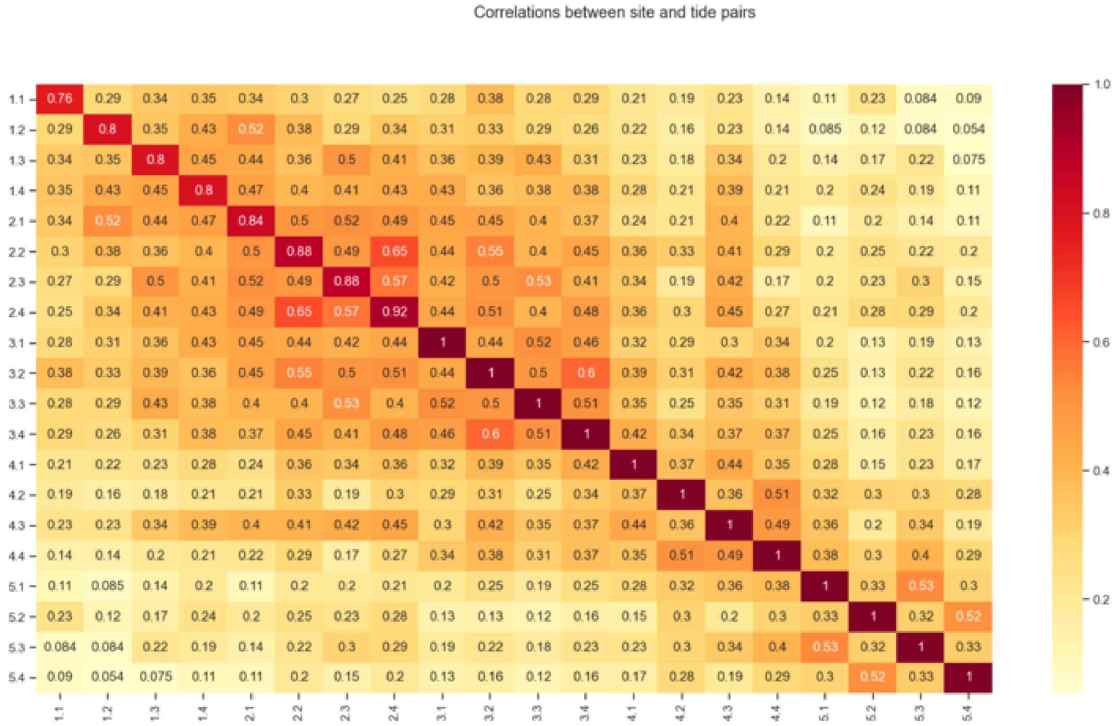


Figure 4.5: Simplified representation of measurement site locations. Dashed line represents site groupings we hypothesise to be similar

non-zero samples, to avoid the noise caused by species with low observation rate. For sites (1, 2 and 4) where there is only data for the initial period, the Pearson correlation was calculated using time-series of length equal to the shortest time-series. A grid showing values between all pairs of sites and tides is shown in figure 4.6. On the axis, a label of $x.y$ represents site x and tide y. On the diagonal axis, we might expect correlation values of 1 between identical time series. However, all abundances are zero for some of the species at certain sites and tides. Library functions will return "NaN" for such cases, and so we adopt the convention of setting the correlation to zero for these. Taking an unbiased sample mean, we divide the sum of correlations by 25, resulting in a non-unitary self-correlation.



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	3.1	3.2	3.3	3.4	4.1	4.2	4.3	4.4	5.1	5.2	5.3	5.4
1.1	0.76	0.29	0.34	0.35	0.34	0.3	0.27	0.25	0.28	0.38	0.28	0.29	0.21	0.19	0.23	0.14	0.11	0.23	0.084	0.09
1.2	0.29	0.8	0.35	0.43	0.52	0.38	0.29	0.34	0.31	0.33	0.29	0.26	0.22	0.16	0.23	0.14	0.085	0.12	0.084	0.054
1.3	0.34	0.35	0.8	0.45	0.44	0.36	0.5	0.41	0.36	0.39	0.43	0.31	0.23	0.18	0.34	0.2	0.14	0.17	0.22	0.075
1.4	0.35	0.43	0.45	0.8	0.47	0.4	0.41	0.43	0.43	0.36	0.38	0.38	0.28	0.21	0.39	0.21	0.2	0.24	0.19	0.11
2.1	0.34	0.52	0.44	0.47	0.84	0.5	0.52	0.49	0.45	0.45	0.4	0.37	0.24	0.21	0.4	0.22	0.11	0.2	0.14	0.11
2.2	0.3	0.38	0.36	0.4	0.5	0.88	0.49	0.65	0.44	0.55	0.4	0.45	0.36	0.33	0.41	0.29	0.2	0.25	0.22	0.2
2.3	0.27	0.29	0.5	0.41	0.52	0.49	0.88	0.57	0.42	0.5	0.53	0.41	0.34	0.19	0.42	0.17	0.2	0.23	0.3	0.15
2.4	0.25	0.34	0.41	0.43	0.49	0.65	0.57	0.92	0.44	0.51	0.4	0.48	0.36	0.3	0.45	0.27	0.21	0.28	0.29	0.2
3.1	0.28	0.31	0.36	0.43	0.45	0.44	0.42	0.44	1	0.44	0.52	0.46	0.32	0.29	0.3	0.34	0.2	0.13	0.19	0.13
3.2	0.38	0.33	0.39	0.36	0.45	0.55	0.5	0.51	0.44	1	0.5	0.6	0.39	0.31	0.42	0.38	0.25	0.13	0.22	0.16
3.3	0.28	0.29	0.43	0.38	0.4	0.4	0.53	0.4	0.52	0.5	1	0.51	0.35	0.25	0.35	0.31	0.19	0.12	0.18	0.12
3.4	0.29	0.26	0.31	0.38	0.37	0.45	0.41	0.48	0.46	0.6	0.51	1	0.42	0.34	0.37	0.37	0.25	0.16	0.23	0.16
4.1	0.21	0.22	0.23	0.28	0.24	0.36	0.34	0.36	0.32	0.39	0.35	0.42	1	0.37	0.44	0.35	0.28	0.15	0.23	0.17
4.2	0.19	0.16	0.18	0.21	0.21	0.33	0.19	0.3	0.29	0.31	0.25	0.34	0.37	1	0.36	0.51	0.32	0.3	0.3	0.28
4.3	0.23	0.23	0.34	0.39	0.4	0.41	0.42	0.45	0.3	0.42	0.35	0.37	0.44	0.36	1	0.49	0.36	0.2	0.34	0.19
4.4	0.14	0.14	0.2	0.21	0.22	0.29	0.17	0.27	0.34	0.38	0.31	0.37	0.35	0.51	0.49	1	0.38	0.3	0.4	0.29
5.1	0.11	0.085	0.14	0.2	0.11	0.2	0.2	0.21	0.2	0.25	0.19	0.25	0.28	0.32	0.36	0.38	1	0.33	0.53	0.3
5.2	0.23	0.12	0.17	0.24	0.2	0.25	0.23	0.28	0.13	0.13	0.12	0.16	0.15	0.3	0.2	0.3	0.33	1	0.32	0.52
5.3	0.084	0.084	0.22	0.19	0.14	0.22	0.3	0.29	0.19	0.22	0.18	0.23	0.23	0.3	0.34	0.4	0.53	0.32	1	0.33
5.4	0.09	0.054	0.075	0.11	0.11	0.2	0.15	0.2	0.13	0.16	0.12	0.16	0.17	0.28	0.19	0.29	0.3	0.52	0.33	1

Figure 4.6: Pearson correlation for all site and tide pairs, averaged over all species

We can see that in general, sites are more correlated with measurements taken at the same tide. Furthermore, we can see that measurements correlate more with measurements at the same type of tide. An ebb tide is when the tide flows away from the shore, and a flood tide is when the water level rises. Therefore, as expected, flood tide measurements correlate more with the other flood tide measurement taken at that site, than with the ebb tide measurements, and vice versa. The sites are also ordered in distance from the sea, with site 1 being closest to the sea, and site 5 being furthest upriver. We notice that there is a higher correlation between ebb measurements at a downstream site and flood measurements at an upstream site, than other pairs. This makes sense as some species may be at a downstream site during an ebb tide and be moved upstream during a flood tide to another site.

We observe that in general sites {1, 2, 3} and {4, 5} seem to show a higher correlation with themselves than with other sites. However, the increase in correlation is relatively small. Therefore, we decided that the correlation between these sites was not significant enough to justify averaging data between sites {1, 2, 3} and {4, 5}, where data exists for all sites. Instead we choose to focus on only sites 3 and 5, which have measurements for the entire length of the observation period.

4.6 Species name linking

The two different datasets also posed a challenge due to them containing inconsistent species. The species abundance dataset contains observations for 146 species, while the food-web consists of only 71 species. The extra species in the abundance dataset are mostly due to them not falling inside the scope of the food-web studies. Many of the species that have been observed are not a typical member of the ecosystem, and may be a rare sighting, with zero abundance at most times, or do not interact significantly with the ecosystem. On the other hand, the food-web also contained some species not recorded in the abundance data. Some species found in fish guts, such as insects like *Thysanoptera* (Thrips), were not aquatic species and hence not included in the abundance sampling. Furthermore, some species in the same genus were difficult to distinguish during direct observation of fish guts, and so are grouped together into a single node in the food-web. Therefore, we cannot say that the species in either dataset are subsets of the other.

Therefore, we manually created a mapping between species in the abundance dataset, to the species in the food-web. This was done by understanding the common names of all species using web resources and matching species in the time series dataset to species on the food-web. The many species that do not have a corresponding food-web entry are not included in the map. For some food-web nodes that encompass multiple species, such as "Crab", multiple species from the time series dataset were mapped. Therefore, the created mapping is neither injective nor surjective.

We found that species also frequently have scientific names that are synonyms. These can occur when a species is discovered independently in multiple locations or when existing taxa definitions change. Unlike other contexts, these synonyms should not be interchangeable since each refers to a particular circumscription, position and rank, and older names should not be accepted. For this investigation, when we have come across species name synonyms, we have assumed them to refer to the same species. Examples of this include:

- *Melicertus.kerathurus* being a synonym to *Penaeus.kerathurus*
- *Chelon.labrosus* to *Liza.chelo*, allowing us to map *Chelon.labrosus* to the "Liza Sp." node in the food-web data

Chapter 5

Convergent Cross Mapping

5.1 Introduction

It is commonly stated that correlation does not imply causation. The reverse of this is also false, since a lack of correlation does necessarily mean a lack of causation. This obviously makes it difficult to infer causation from correlation. A well known test for causality that overcomes this problem is the Granger causality test, which as described in the background, can measure causal lag and strength by decomposing the cross spectrum between two variables[35].

However, a core assumption of the Granger causality test is the separability of the system. This creates problems in a dynamical system with weaker coupling, such as ecosystems which have been shown to have weak to intermediate interaction strengths between species[36] and therefore must be considered in their entirety. Sugihara et al. demonstrates that three main Granger causality methods (vector autoregression, conditional mutual information and spectral decomposition) fail to consistently identify causality when applied to ecosystems, and the separability assumption is violated[8]. In order to provide an alternative causality test for this class of system they introduce the convergent cross mapping (CCM) test for detecting causality, which we will explain, starting from the core concepts, and then apply to the Guadalquivir estuary ecosystem.

5.2 Convergent Cross Mapping

The phase space of a system represents all possible states of a system, which in the case of our ecosystem is the abundances of all species. Assuming the system to be a continuous-time dynamical system, the phase space can then be assumed to be a smooth manifold (a topological space with Euclidean properties locally around any point) of dimension m[37]. The values of this system will tend to evolve towards a set of values called the attractor manifold. Once the values of the system are on the attractor, the trajectory of the system must remain on the attractor forward in time. In 1944, Hassler Whitney proved that manifolds can be regarded as subspaces of some Euclidean space[38]. The theorem states that every m dimensional differentiable manifold can be embedded in R^{2m}.

Based on Whitney's embedding theorem, Floris Takens then developed his delay embedding theorem in 1981[39]. The idea behind Takens' embedding theorem is that information about a system is retained in a one-dimensional observation or measurement of the system. A one-dimensional observation with the estuary ecosystem as the dynamical system under consideration can be the abundance of a single species. The abundance of this species is assumed to be following a trajectory defined by the attractor manifold. We create a function called a delay coordinate map by taking multiple, uniformly spaced, time-lagged values of the system. If the one-dimensional time series is $X(t)$ and the chosen lag is τ for E delays, we sample

$$\{X(t), X(t-\tau), X(t-2\tau), \dots, X(t-(E-1)\tau)\}$$

By Takens' delay embedding theorem, this delay coordinate mapping is then a diffeomorphism (bijective mapping between manifolds) of the attractor of the original system[40]. Furthermore, the number of delays, E, required, is at most $2m+1$, where m is the dimension of the attractor manifold, and often less as demonstrated later in Section 1.3 for the Lorenz system.

If two time-series variables, X and Y are causally linked, they will share a common attractor manifold[39]. The signature of X will be detectable in Y, allowing us to recover information about Y from the time-series of X, and vice-versa. With time series X and Y, let us construct two time delayed embeddings M_x and M_y, called the shadow manifolds. If X and Y are casually linked and share the same attractor manifold, then M_x and M_y should also be embeddings of the same attractor manifold. We can test this by seeing if nearby points in M_x correspond to nearby points in M_y and nearby points in M_y correspond to nearby points in M_x. We do so by using a nearest-neighbour algorithm called simplex projection. As the time-series lengths increase, more points on the attractor manifold are filled in and the estimates of one variable using the other improve, resulting in a convergence in estimation skill. This process is called **convergent cross mapping** (CCM)[8].

If instead, one variable, X, drives the other variable, Y, then information about X can be recovered from the values of Y, but information about Y can not be found in X. This will mean that M_y can be used to estimate points on M_x, but not conversely. In the case of the ecosystem, this will suggest a one-directional interaction, with the population of one species driving the population of the other.

5.3 Lorenz System

We will demonstrate time-delay embedding and CCM on the Lorenz system, a famous chaotic system given by the three ordinary differential equations known as the Lorenz equations[41]:

$$\frac{dx}{dt} = \sigma(y - x)$$
$$\frac{dy}{dt} = x(\rho - z) - y$$
$$\frac{dz}{dt} = xy - \beta z$$

We shall use the original values $\sigma = 10, \beta = \frac{8}{3}, \rho = 28$. With these values we calculate 100 time-steps, using initial values of $\{1, 1, 1\}$, to get figure 5.1.

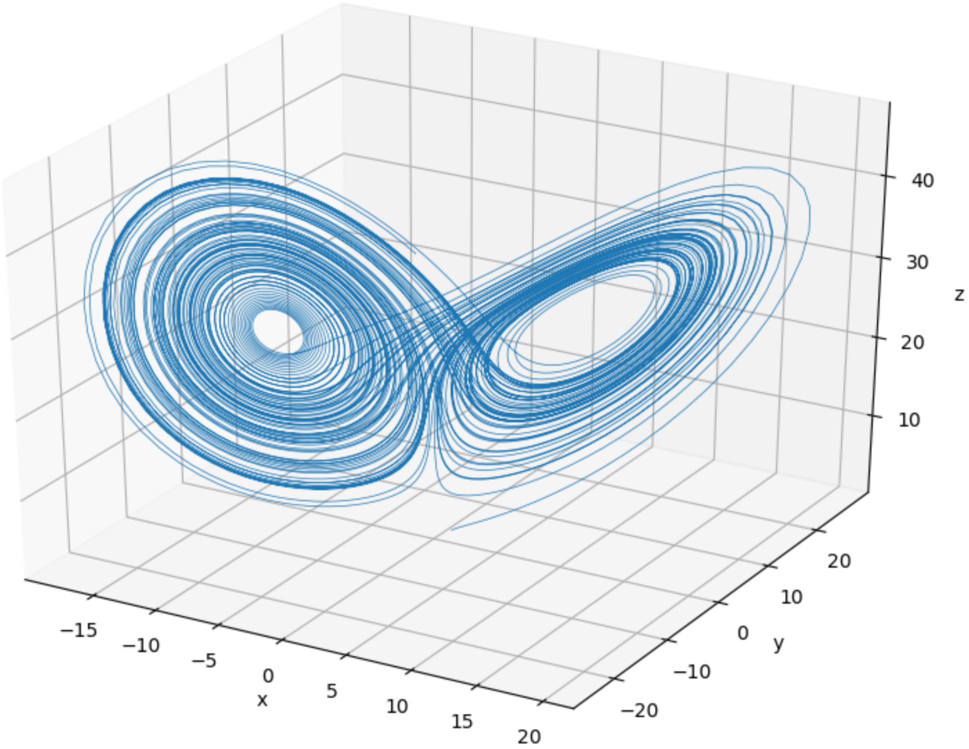


Figure 5.1: Visualisation of 100 time-steps of the Lorenz system, using matplotlib

We can see states of this system follow a butterfly-shaped attractor manifold, which in this case is a strange attractor due to its fractal structure. We now create a time-delay embedding for the variable y, taking an embedding dimension of $E = 3$ and lag values $\tau = 1$. Plotting each lagged time-series on an axis, we get the plot shown in figure 5.2. There exists a diffeomorphism between the original manifold and the shadow manifold in figure 5.2, meaning there is a one-to-one mapping between them, and the shadow manifold preserves all the topological properties of the attractor manifold.

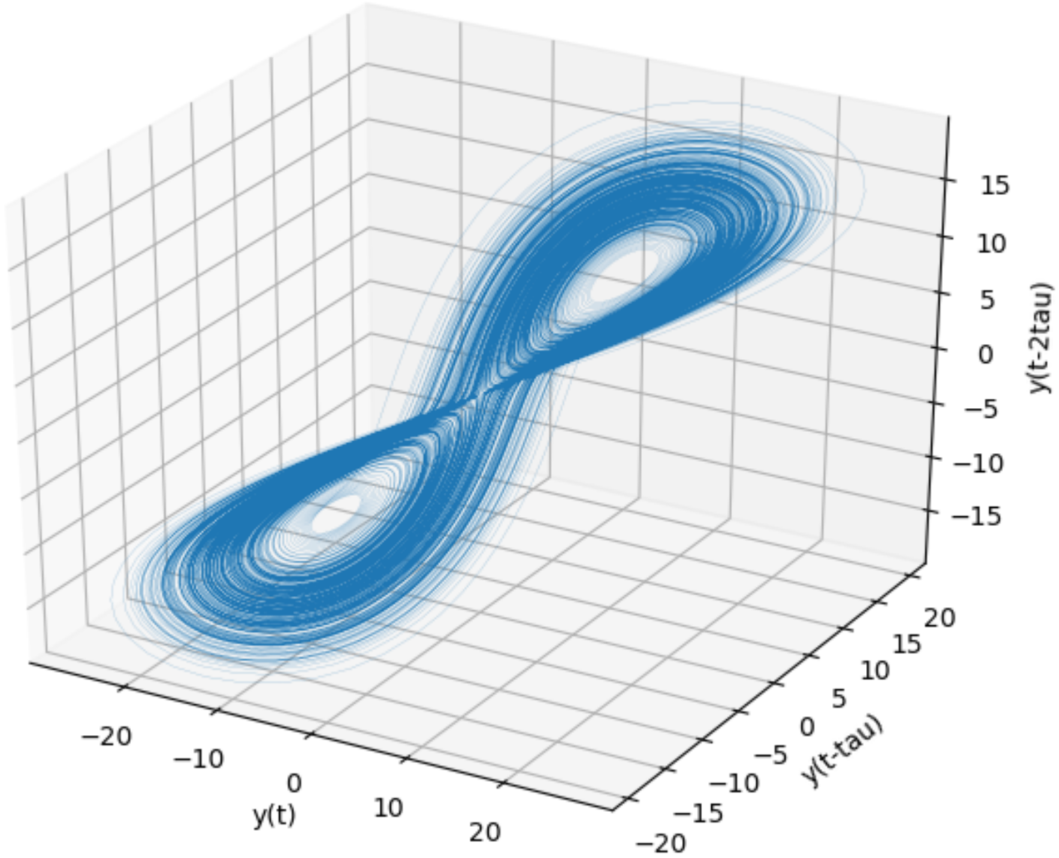


Figure 5.2: Time delay embedding of Lorenz system y variable

After creating shadow manifolds for the other variables, we can run the CCM method. We shall do this using the pyEDM Python library, details of which are later explained. The three plots in figure 5.3 show the cross-mapping between each pair of variables. The y-axis shows the Pearson correlation of the estimated nearby points obtained from the simplex projection algorithm. We can see that on the first plot, the estimation in both directions increases in accuracy, with the correlation converging to near 1. In the other two plots, the cross-maps for x cross-map z and y cross-map z show an increase but the converse is false. This suggests that information about z can be recovered from x and y and that z drives these two variables. We know from the Lorenz equations that this is true, but also that x and y should drive z. The reason that the CCM estimate does not increase is because the Lorenz attractor is one of several special cases that does not satisfy the requirements for manifold reconstruction. The lobes of the Lorenz attractor manifold are symmetric with respect to z. This leads to the two fixed points on the attractor manifold having the same value of z. These correspond to a single fixed point on the shadow manifold of z, and lead to the system dynamics being impossible to reproduce[42]. We do not worry about this special case during our CCM usage since any noise in a real-world example, like an ecosystem, will lead to the symmetry being broken.

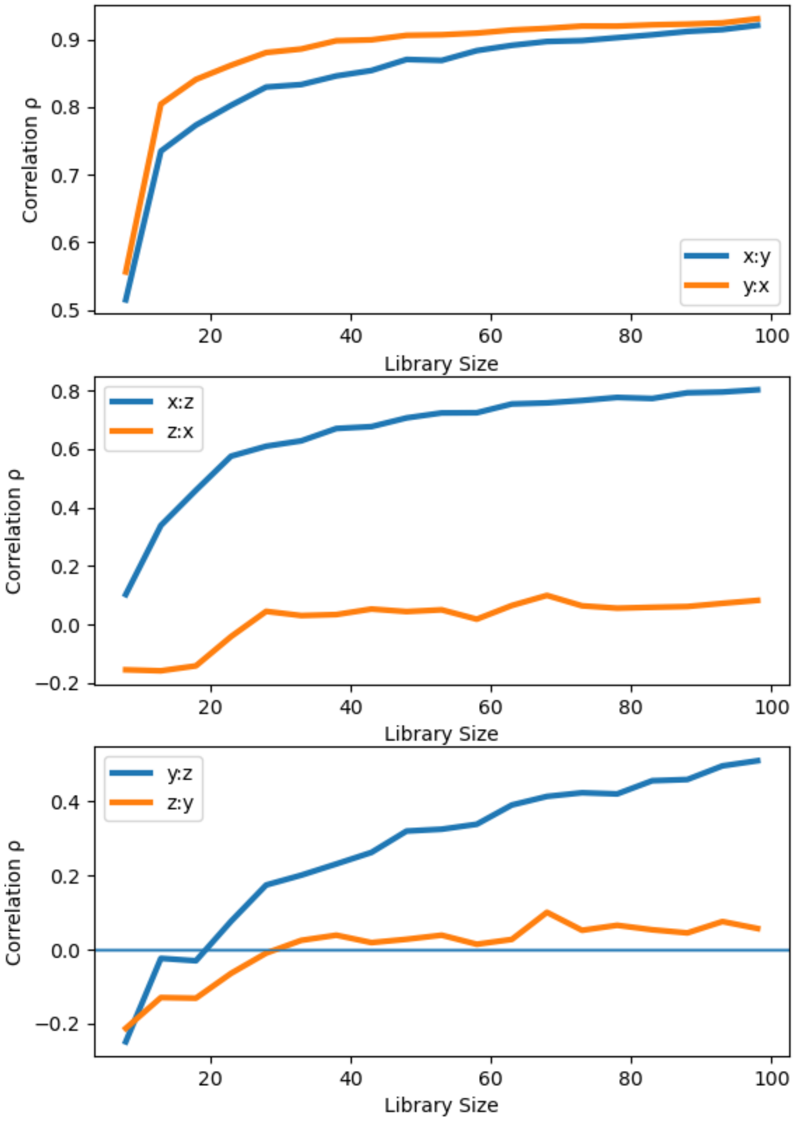


Figure 5.3: CCM on the Lorenz system using pyEDM, with a lag of $\tau = 1$ and embedding dimension of $E = 3$

5.4 PyEDM

The Sugihara Lab has created the pyEDM Python package, implementing the CCM method introduced in Sugihara et al.[8], as well as other nonlinear dynamical system modelling methods[43]. We also considered the skccm Python library[44], but chose pyEDM as it has been more extensively reviewed, and is used in several publications, including the original paper. In order to understand, test and demonstrate the usage of the library, we replicated several of the CCM results found in the original publication[8].

The study by Sugihara et al. demonstrates the CCM method by testing for causality between anchovies, sardines and sea-surface temperature at two locations in California[8]. There are competing hypotheses explaining the pattern of alternating dominance that is often found between sardine and anchovy populations, which Sugihara et al. aim to address. Using the same publicly available dataset of sardine and anchovy populations, and sea-surface temperature measurements, we can replicate the results from the study. Doing so produces the CCM prediction score plots shown in figure 5.4.

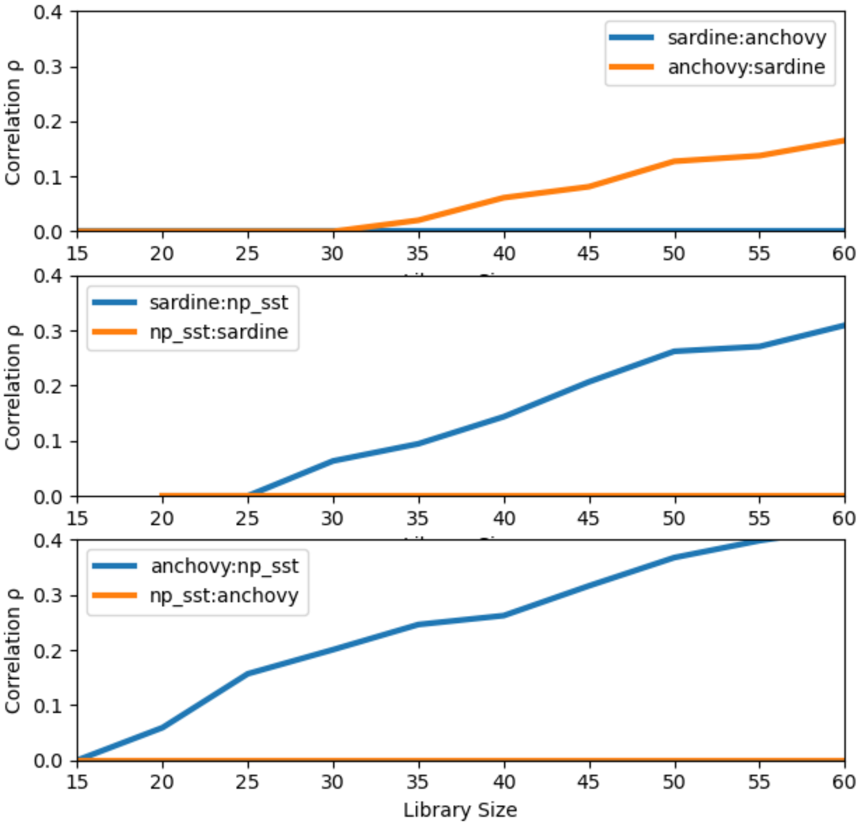


Figure 5.4: Output of pyEDM CCM on anchovy and sardine populations, and sea surface temperature, with an embedding dimension of 3, and time delays of 2, -5 and 1

With these results, we can draw similar conclusions to Sugihara et al. The anchovy and sardine cross-maps show a small amount of cross-map signal. However, both sardines and anchovies show strong cross-map signal against sea surface temperatures. This suggests that there is a detectable signature from the sea surface temperature in the sardine and anchovy time series, and so their populations are primarily both driven by the temperature.

It is important to highlight the time delay used in these CCMs was 2, -5 and 1. Original CCM with a default time delay of -1 showed much weaker cross-map signals. Details on how these time delay values were chosen are explained in the optimal lag Section 5.6.

The interactions of Didinium and Paramecium, a classic predator-prey system, were similarly tested using CCM, as in Sugihara et al[8]. The results are shown in figure 5.5. There is significant CCM in both directions. However, the stronger signal between Paramecium (the prey) cross-mapped with Didinium (the predator) suggests that the Didinium has a stronger driving force on the Paramecium, and as such this is a top-down ecosystem[8].

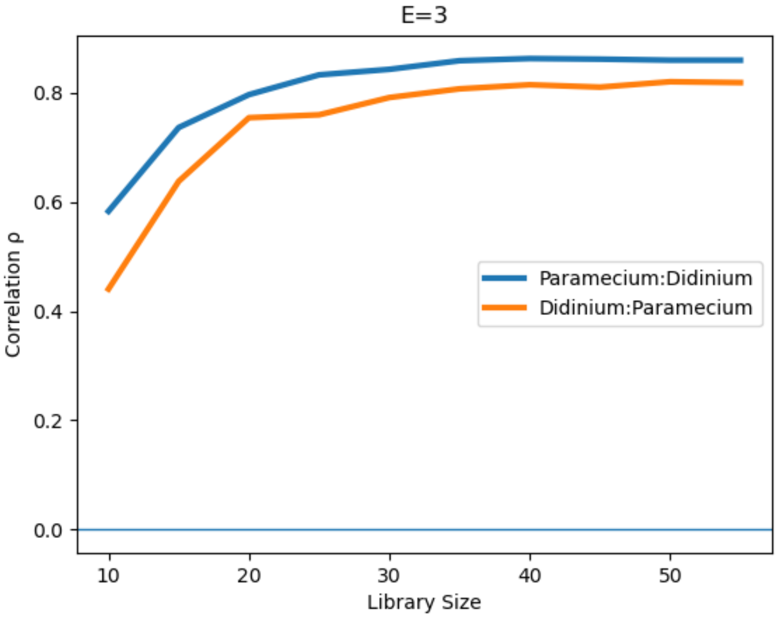


Figure 5.5: Outputs of pyEDM CCM between Paramecium and Didinium populations, with an embeddng dimension of 3 and time-delay of 1

5.5 Optimal Dimension

As explained before, a sufficient embedding dimension for CCM has been shown to be $E > 2m$. However, this is not a necessary condition, and in practical cases, a smaller dimension can often be used. For example, Takens' theorem states that a sufficient condition for the Lorenz system with an attractor manifold dimension of $m = 3$ is $E = 2m + 1 = 7$. However, as shown in the example, a diffeomorphic shadow manifold (for two of the variables) can be constructed with an embedding dimension of 3. When constructing a shadow manifold, we want the trajectory of variables to be deterministic and only depend on their current position. This is violated whenever there is an intersection in the embedding. We avoid this by using sufficient dimensions for the embedding. Generically, a dimension of $2m + 1$ is needed, but uniqueness can often be preserved with lower dimension embeddings[45].

Finding a good embedding dimension in practice is often done empirically. Multiple embedding dimensions can be tried, and the dimension which provides the best cross-map estimates accepted. pyEDM function contains the *EmbedDimension* function which runs CCM on a range of dimensions, up to a specified maximum dimension. Using this, we attempted to determine an optimal embedding dimension that could be used for all CCMs of the ecosystem. We used the library function to run CCM on all pairs of the top 25 species with embedding dimensions between 1 and 10. The optimal dimension was then chosen for each case by selecting the embedding dimension that gave the highest final correlation score. These optimal dimensions can be seen plotted on a histogram in figure 5.6. The first histogram shown in the top left of the figure shows no clear trend for an optimal embedding dimension. However, this is due to many of the pairs of species not being causally linked, and hence scoring a low CCM score regardless of the embedding dimension. Therefore, we filter out these cases in the other histograms, by only recording embedding dimensions for CCMs that score above a threshold score, for thresholds between 0.1 and 0.5. We can see that with threshold scores of 0.3 or 0.4, which imply that there is significant causality between the species, an optimal embedding dimension lies around $E = 3$ for many of the pairs. As such, we choose this value as an embedding dimension to use for our initial prediciton model.

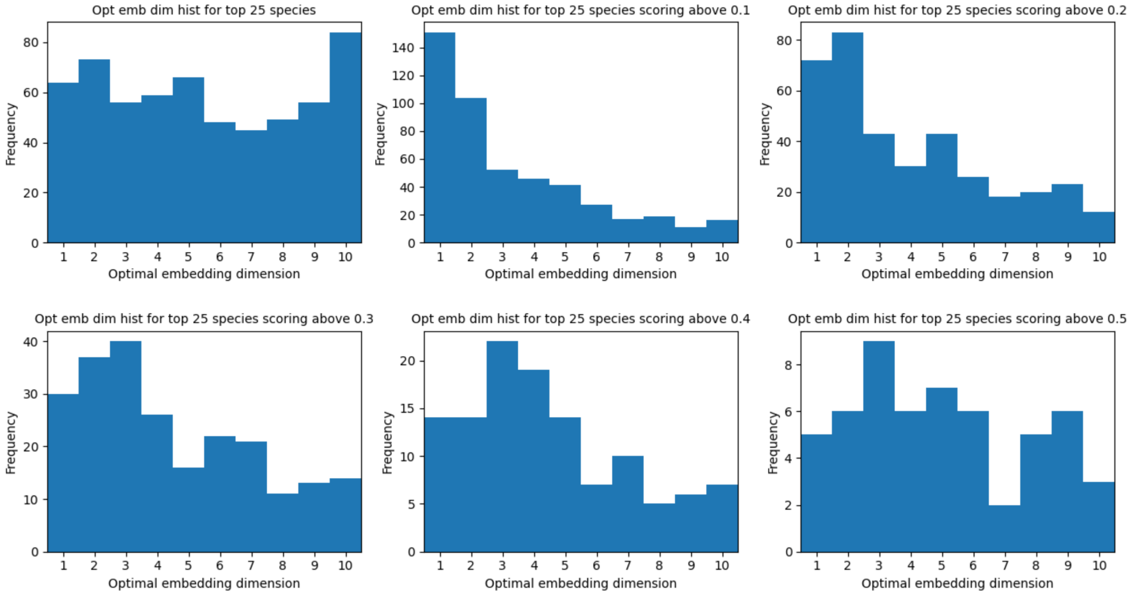


Figure 5.6: Optimal embedding dimensions for all pairs of the top 25 species. All embeddings used a delay of $\tau = 1$.

5.6 Optimal Lag

When reconstructing the attractor of our system using the delay embedding theorem, we take E values

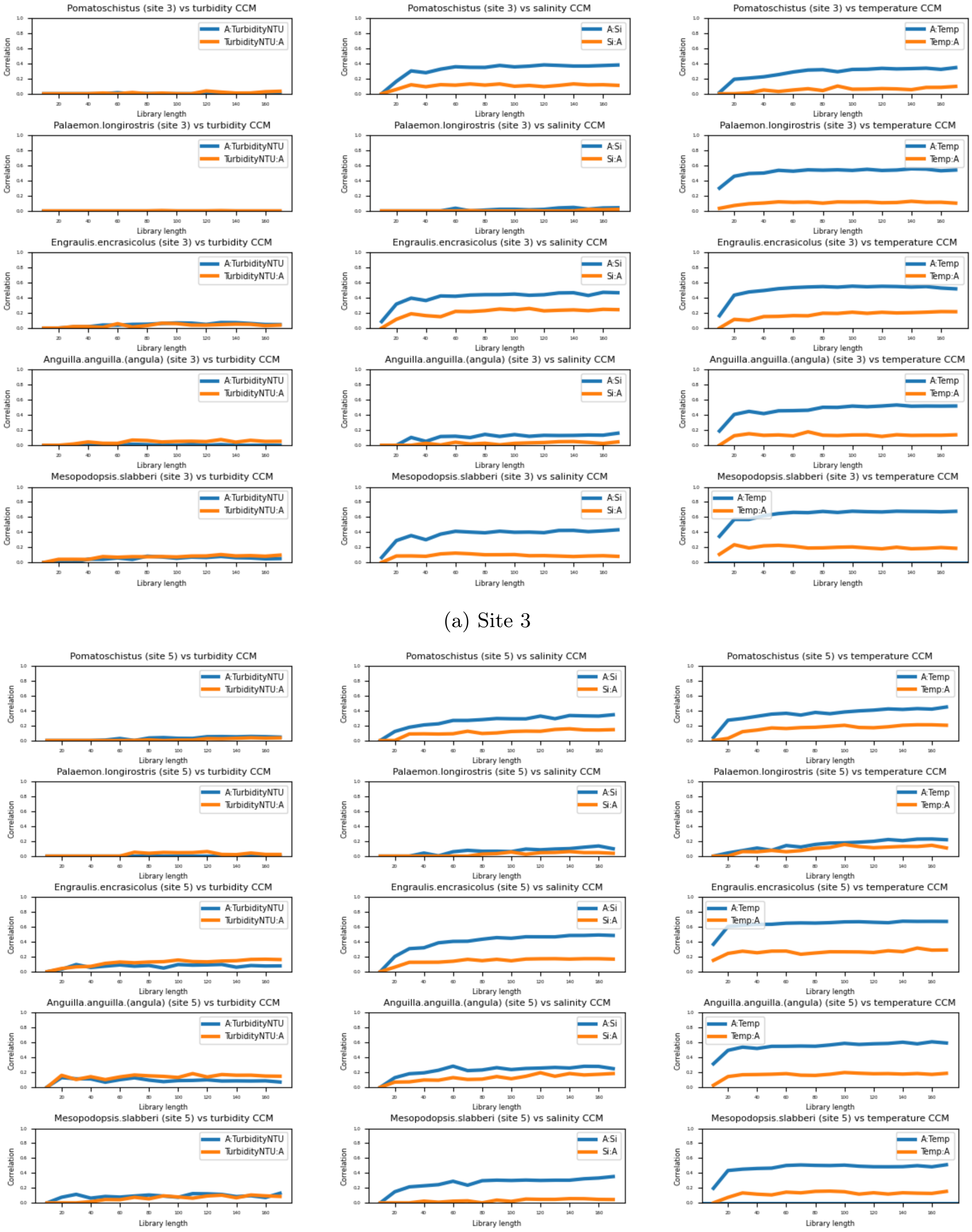
$$X_t, X_{t+\tau}, X_{t+2\tau} \cdots$$

where E is the embedding dimension used. τ is the time lag between the points used to reconstruct the attractor. The delay embedding theorem states that the value of τ is not important. However, in practice, if the value of τ is too small, the observed values will be too correlated and cause the attractor to cluster around the diagonal in the embedding space. If τ is too large, the observed values will be too uncorrelated and cause the attractor to become obscure[46].

As such, in an optimised model, we can empirically find the optimal lag by repeating the attractor reconstruction using multiple values of τ, and accept the τ that provides the highest prediction score. This is done by running the pyEDM CCM algorithm over a range of τ from -5 to 5, not including 0 as we require different lagged points. The final (terminal) estimate score of the time-series is taken as the score for that value of τ. The results of this optimisation are shown later in table 5.2.

5.7 Environmental Influence on Species

We start by attempting to determine if there is any causal link between the species abundances and the environmental conditions. We run CCM for the 5 most dominant species, with respect to the fraction of non-zero measurements, against the observed water salinity, turbidity and temperature, which have been recorded alongside the abundance at each measurement. We choose to separate the data by site, since the environmental conditions are likely to vary between sites. Since sites 3 and 5 have observations for the full length of the observation period, we will focus on them.



(a) Site 3

(b) Site 5

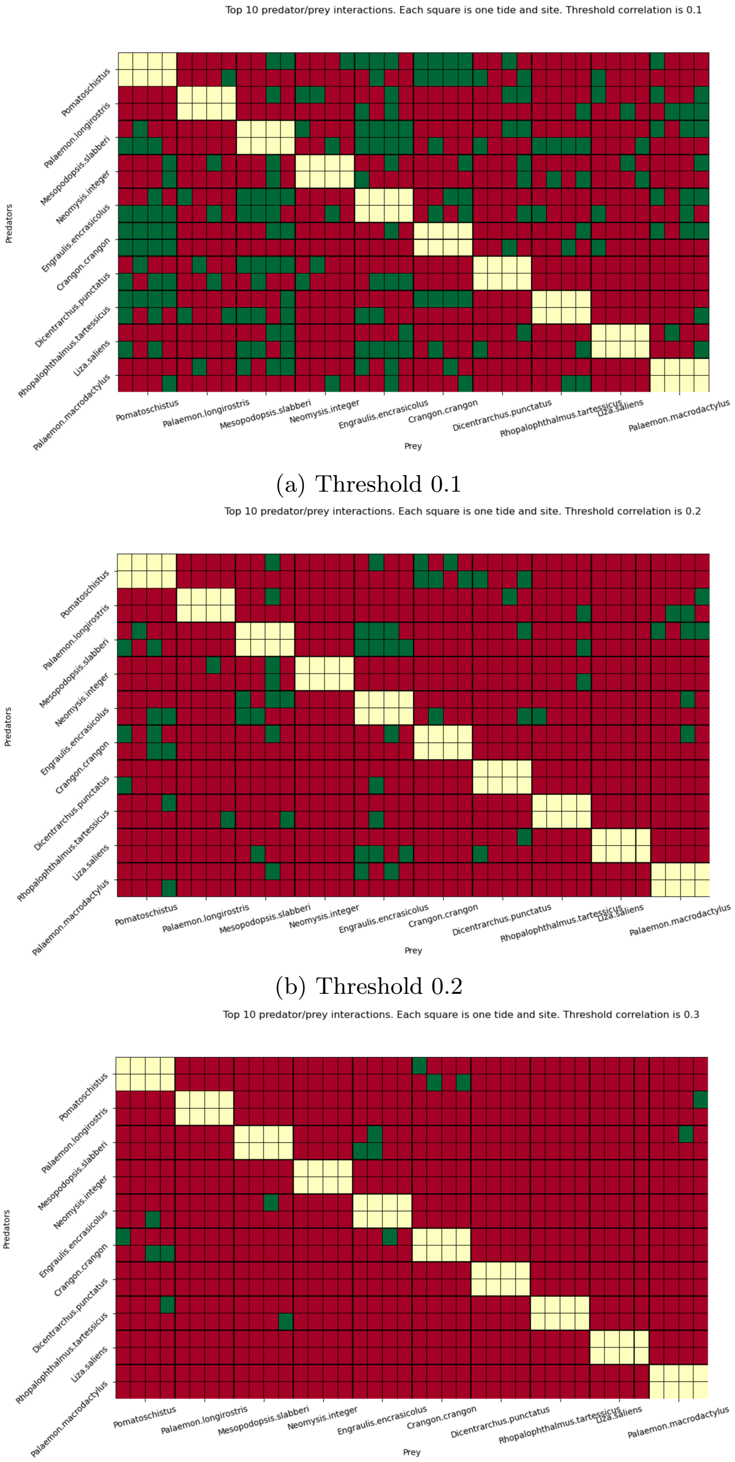
Figure 5.7: Results of top 5 species CCM with salinity, turbidity and temperature at sites 3 and 5

In all the plots in figure 5.7, the blue line represents the influence of the environmental variable on the species abundance, and the orange line represents the converse. We see that at either site, the turbidity has very little influence on species abundance. However, salinity appears to influence some species more than others. We find that 3 out of 5 species converge to a cross-map estimate correlation above 0.4. This suggests that water salinity is a driving force in the abundance of some, but not all, species. Temperature shows the strongest causality towards species abundance, with all species showing significant cross-map estimate correlation. Cross-estimates of the temperature shadow manifold using the *Mesopodopsis slabberi* shadow manifold score a correlation above 0.7. This can potentially be explained by temperature having a strong influence on species life-cycles, with most species being more abundant in certain months of the year, as shown in figure 4.4.

5.8 CCM Evaluation

Since we have observed the environmental conditions to have a significant influence in determining species abundance, we decided to perform the convergent cross mapping between species at individual sites and tides, rather than combining the data from all sites and tides into one mapping. This will allow for less noise in the data, due to environmental factors such as temperature or salinity. It leads to the data being evenly spaced out, with one observation each month, rather than measurements on four consecutive tides each month. Keeping the sites and tides separate also allows us to have 8 different CCM models, using combinations of the 4 tides and 2 sites. We will be able to run these multiple models as an ensemble method, where the overall output is based on what the majority of individual models predict. Each site and tide has approximately 200 measurements, which has been shown to be sufficient for successful cross mapping prediction[8].

We ran the CCM algorithm between all combinations of the top 10 species at each site and tide. We then observed the terminal value of the cross mapping prediction score to determine how well the abundance of one species can predict the other. This was used to produce the grids shown in figure 5.8, where each square corresponds to the CCM between two species at a specific site and tide. Squares are coloured green if the prediction score exceeds a specified threshold value. The ensemble of CCM models can then vote on a final prediction on whether there is causality between each pair of species, based on a specified required amount of CCM models predicting an interaction.



(a) Threshold 0.1

(b) Threshold 0.2

(c) Threshold 0.3

Figure 5.8: CCM results of top 10 species done at each site and tide. Green signifies terminal correlation above threshold value. Red signifies terminal correlation below threshold value.

We can see from the grids that, as expected, a higher threshold value results in fewer green squares.

Using the number of green squares for each species pair, we can now make a prediction on whether there is an interaction between the two species. We do this for a range between 1 and 5 of required green squares for an interaction to be predicted. We can then evaluate the prediction by using the species name map to check if there is an actual interaction between the two species in the food-web. We also confirm that the species inputted to the model exist in the food-web, and throw an error if they do not.

To analyse the results, we split the prediction into different totals:

- True positives (TP) - the number of interactions correctly predicted.

- False positives (FP) - the number of interactions predicted that do not actually exist.

- True negatives (TN) - the number of interactions that do not exist correctly predicted.

- False negatives (FN) - the number of interactions predicted to not exist that do actually exist.

Using these, we can calculate metrics to evaluate the model performance. Metrics we use are

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Using these metrics allows us to measure how reliably we can trust a positive model prediction (precision) and how many of the interactions the model manages to predict (recall). We use the F1 score to strike a balance between these metrics for an overall score. It is important to observe both precision and recall to reach a reliable model evaluation, since for example, a model that always predicts an interaction will achieve a high recall, but a low precision.

The results for the initial CCM model using an embedding dimension of $E = 3$ and $\tau = 1$ for all CCMs are shown in table 5.1, evaluated against a directed food-web. We can see that the model performs best, in terms of F1 score, with a threshold value of 0.1 and only 1 green square required to predict an interaction, giving an F1 score of **0.569**. With these parameters, the recall is **0.892**, meaning that 89.2% of interactions between the species are predicted. However, the precision is rather low with a score of **0.418**, meaning that only 41.8% of predicted interactions are actually correct. This shows that the model is over-predicting the number of interactions, with a high number of false positives. Increasing the threshold and required majority of green squares, as expected, improves the precision but reduces the recall as predictions now face stricter criterion. The model using a threshold of 0.2 and required majority of 1 green square provides a more balanced model, with a higher precision but lower recall, and a F1 score only slightly lower of **0.565**.

Threshold **0.1**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	33	4	0.418	**0.892**	**0.569**
	46	17			
2	25	12	0.431	0.676	0.526
	33	30			
3	17	20	0.486	0.459	0.472
	18	45			
4	12	25	0.462	0.324	0.381
	14	49			
5	7	30	**0.500**	0.189	0.275
	7	56			

Threshold **0.2**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	24	13	**0.500**	0.649	0.565
	24	39			
2	6	31	0.429	0.162	0.235
	8	55			
3	5	32	0.455	0.135	0.208
	6	57			
4	4	33	0.444	0.108	0.174
	5	58			
5	2	35	0.286	0.054	0.091
	5	58			

Threshold **0.3**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	9	28	**0.500**	0.243	0.327
	9	54			
2	3	34	0.375	0.081	0.133
	5	58			
3	2	35	0.333	0.054	0.093
	4	59			
4	1	36	0.200	0.027	0.048
	4	59			
5	0	37	0.000	0.000	-
	2	61			

Table 5.1: CCM performance on top 10 species, with $E = 3$ and $\tau = 1$ on a directed food-web

As an attempt to improve the model, we tried using optimal values for the embedding dimension and lag in each CCM. This was done by running the CCM for each pair on multiple lags between -5 and 5 and using the lag that gave the highest terminal correlation score. The CCM was then run with embedding dimensions between 1 and 10, with the best embedding dimension being used. The goal here was to reduce the number of false negatives, where the model may fail to predict an interaction because it is badly parameterized for the specific pair of species. The disadvantage of this model is that it is much slower since each pair of species requires 20 CCM runs per site and tide. The results of this optimized model are shown in table 5.2. As expected, the number of false negatives decreases and we achieve a higher recall in all cases, with a best recall of **0.919**. However the best F1 score is **0.548**, a lower score due to the slight reduction in precision for most parameters, due to more false positives. Nevertheless, a higher maximum precision score is achieved of **0.542** with a model using a threshold of 0.1 and required majority of 5. In this case though, the recall is significantly lower at **0.351**.

We find that these results somewhat support the delay embedding theorem in stating that the value of τ is not important, while within a reasonable range. For our dataset, even with the a minimum τ value of $|-1|$, the time between observed values is 1 month, allowing for sufficient distance between points. When we choose an optimized value between -5 and 5, the change in performance is not that substantial.

Threshold **0.1**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	34	3	0.391	**0.919**	**0.548**
	53	10			
2	30	7	0.405	0.811	0.541
	44	19			
3	25	12	0.417	0.676	0.515
	35	28			
4	15	22	0.455	0.405	0.429
	18	45			
5	13	24	**0.542**	0.351	0.426
	11	52			

Threshold **0.2**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	27	10	0.422	0.730	0.535
	37	26			
2	17	20	0.472	0.459	0.466
	19	44			
3	10	27	0.476	0.270	0.345
	11	52			
4	5	32	0.455	0.135	0.208
	6	57			
5	4	33	0.444	0.108	0.174
	5	58			

Threshold **0.3**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	12	25	0.429	0.324	0.369
	16	47			
2	6	31	0.545	0.162	0.250
	5	58			
3	5	32	0.500	0.135	0.213
	5	58			
4	4	33	0.500	0.108	0.178
	4	59			
5	1	36	0.333	0.027	0.050
	2	61			

Table 5.2: CCM performance on top 10 species, using optimal dimension and lag for each pair

5.9 Undirected CCM Evaluation

In general, the main problem with both models is a low precision score due to a high number of false positives. There is one main issue that we can suggest to be the cause of these false positives. One species may drive the population of another species so strongly that they experience a phenomenon called "synchronization". If dynamics of one species becomes "enslaved" to a driving species, it will become an observation function of the driving species, and CCM will operate in both directions, even though it is a unidirectional interaction. Results from evaluating the model against an undirected food-web are shown in figure 5.3. In this case, synchronization is not an issue since all that is required is for the model to identify at least one direction of interaction to predict an undirected food-web edge. On an undirected network the model seems to perform better, with a maximum F1 score of **0.816**. However, this score must be treated with caution since in this case the model is predicting interactions for every single case. This obviously creates a perfect recall score of **1.0**, and the precision of **0.689** seems relatively high due to most of the possible edges (68.9%) being realized. The converse of this occurs for most of the parameters with a threshold of 0.3, with the model achieving a perfect precision score, mostly due to the fact that it only predicts a few interactions that it is certain on. Nevertheless, the model seems to perform better in the undirected case, and can provide balanced predictions with parameters such as a threshold of 0.2

and a required majority of 1. This is due to the weaker prediction requirements (no direction for the interaction needed). However, under these requirements, the model can be viewed as more reliable since it won't face issues due to synchronization between species.

Threshold **0.1**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	31	0	0.689	**1.000**	**0.816**
	14	0			
2	28	3	0.683	0.903	0.778
	13	1			
3	23	8	0.719	0.742	0.730
	9	5			
4	20	11	0.769	0.645	0.702
	6	8			
5	15	16	0.714	0.484	0.577
	6	8			

Threshold **0.2**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	23	8	0.719	0.742	0.730
	9	5			
2	14	17	0.700	0.452	0.549
	6	8			
3	6	25	0.667	0.194	0.300
	3	11			
4	4	27	0.800	0.129	0.222
	1	13			
5	2	29	1.000	0.065	0.121
	0	14			

Threshold **0.3**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	12	19	0.857	0.387	0.533
	2	12			
2	3	28	**1.000**	0.097	0.176
	0	14			
3	2	29	1.000	0.065	0.121
	0	14			
4	2	29	1.000	0.065	0.121
	0	14			
5	2	29	1.000	0.065	0.121
	0	14			

Table 5.3: CCM performance on top 10 species, with $E = 3$ and $\tau = 1$ on an undirected food-web

5.10 Phase-lock Twin Surrogate Method

An alternative method for dealing with false positives due to synchronization is to use the phase-lock twin surrogate method introduced by Ushio et al.[47]. The synchronization is often driven by strong seasonality in the data. The phase-lock twin surrogate method aims to create alternative (surrogate) time-series, which have the same seasonality as the original data, but are not causally linked. The cross-map scores produced by CCM on the surrogate time-series can then be compared to the cross-map results for the original time-series. If time-series that aren't causally linked appear to produce cross-map estimates as well as the original time-series, then we can conclude that the cross-map scores in the original time-series are due to the seasonality driven synchronization.

The first step is to generate the phase-locked twin surrogate time series. This is done by first constructing a recurrence matrix: a matrix that identifies where trajectories in a state space come close together. We create a time-delay embedding of the original time-series, choosing to use an embedding dimension of 3. Let the states of this embedding at time t be represented by the vector $\underline{x}(i)$. The recurrence matrix, $R_{i,j}$, is then constructed using the equation

$$R_{i,j} = \Theta(\delta - ||\underline{x}(i) - \underline{x}(j)||)$$

where Θ is the Heaviside function and δ is a specified parameter. This results in a binary matrix, where a 1 represents the states at time i and j being closer than δ to each other, in terms of Euclidean distance. In practice, the best method we found for creating this recurrence matrix was by using library functions to:

1. create a distance matrix of distances between all states
2. calculate a specified quantile value of all the distances
3. set all elements below the quantile value to 1, and all elements above the quantile value to 0

Using a quantile value rather than a δ leads to a consistent number of "1"s in the matrix. Studies have found that having between 5% and 20% of the recurrence matrix as "1"s leads to good surrogate time-series[47].

Next, twin points on the recurrence matrix are identified. Twin points are points which share the same neighbourhood, and so cannot be distinguished when only their neighbourhoods are considered. On the recurrence matrix, this happens when two columns are identical:

$$R_{k,i} = R_{k,j} \forall k$$

The surrogate time-series can now be constructed. A random point on the original time-series is chosen as the starting point of the surrogate time-series. We also add a constraint for the starting point of the surrogate time-series to be the same month as the starting point of the original time-series, for the seasonality to match. We then add points to the surrogate time-series by following the original time-series until we reach a point that has one or more twins. We then, with equal probability, choose to jump to one of these twin points, or to remain at the current point. To maintain seasonality, we also add here, the constraint that to jump to a point, it must be the same month. The surrogate time-series then continues to be built in this style until it is the same length as the original time-series.

The result of this is a time-series that appears to have the same seasonality as the original time-series, but is not causally linked. Generated phase-lock twin surrogate time-series for the 5 most dominant species are shown in figure 5.9.

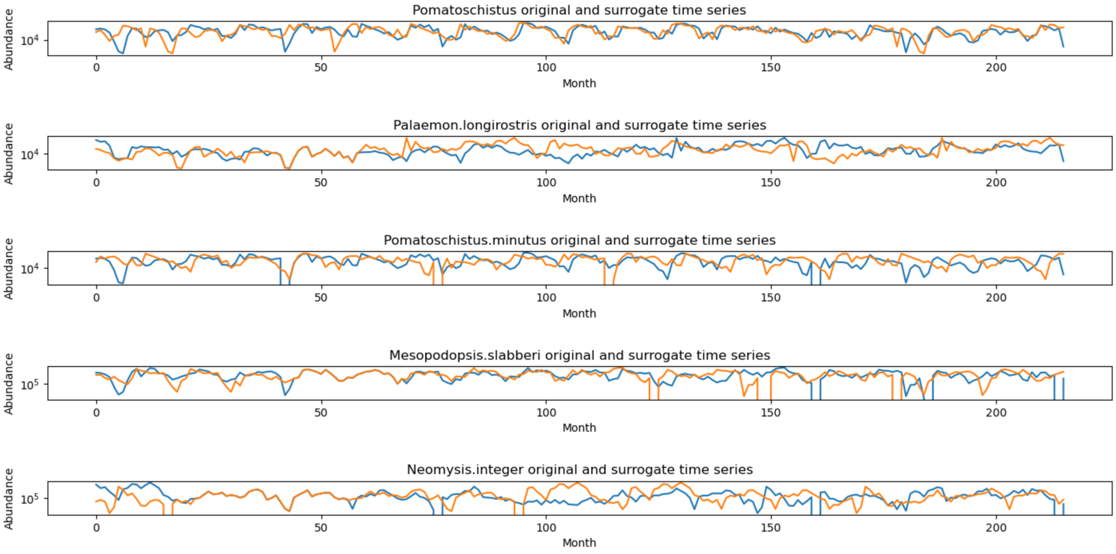


Figure 5.9: Phase-locked twin surrogate time-series for 5 most dominant species. Blue lines represent the original time-series and orange lines represent the surrogate time-series.

Threshold **0.1**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	13	20	0.448	**0.394**	**0.419**
	16	41			
2	6	27	0.545	0.182	0.273
	5	52			
3	5	28	0.833	0.152	0.256
	1	56			
4	3	30	0.750	0.091	0.162
	1	56			
5	3	30	**1.000**	0.091	0.167
	0	57			

Threshold **0.2**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	11	22	0.500	0.333	0.400
	11	46			
2	4	29	0.500	0.121	0.195
	4	53			
3	1	32	1.000	0.030	0.059
	0	57			
4	1	32	1.000	0.030	0.059
	0	57			
5	0	33	0.000	0.000	0.000
	0	57			

Threshold **0.3**

Required majority	TP/FP	FN/TN	Precision	Recall	F1
1	5	28	0.556	0.152	0.238
	4	53			
2	1	32	1.000	0.030	0.059
	0	57			
3	1	32	1.000	0.030	0.059
	0	57			
4	0	33	0.000	0.000	0.000
	0	57			
5	0	33	0.000	0.000	0.000
	0	57			

Table 5.4: CCM performance on top 10 species with optimal E, τ and the phase-lock twin surrogate condition on a directed food-web

We create phase-locked twin surrogate time-series for every species and run the CCM method between all pairs. We can then impose a condition on the model, requiring the final correlation score between the original time-series to be higher than the final correlation score of the CCM between the surrogate time-series. The results for when this condition is applied are shown in table 5.4.

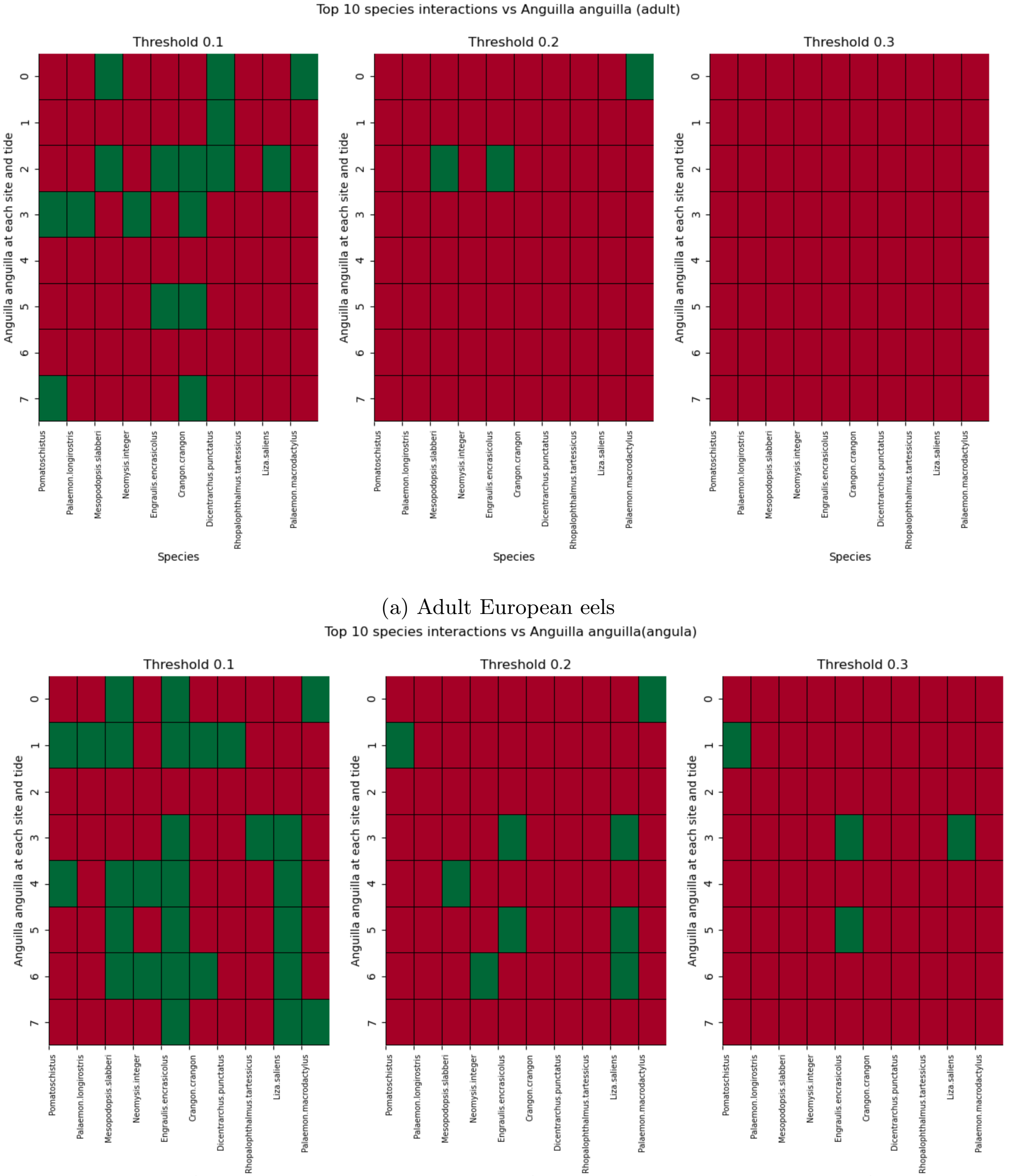
We find that the model does produce better precision scores, with precisions for some parameters being 100%. However, the F1 scores are generally much lower due to the lower recall scores. With the phase-lock twin surrogate condition, the model predicts much fewer positives, leading to more false negatives and lower recall. The model could potentially be improved by adjusting the twin-surrogate time series construction to differ even more from the original time-series, for example by allowing larger jumps between points with a greater δ parameter. This model may still be considered useful if the objective is to obtain predictions that are made with a high degree of certainty. In general, this is often useful for making decisions that carry a high-level of risk, but in the case of reconstructing a food-web, we do not value precision much more than we value recall.

5.11 European Eel Investigation

Anguilla anguilla, commonly known as the European eel, undertake a spawning migration over 5000km from Europe to the Saragasso Sea[48]. The details of this migration are still not fully understood by scientists. Through our communications with Cesar Vilas[2], we have learnt that the European Eel spend the majority of their lives in fresh water, including upstream in the Guadalquivir river. When they become sexually mature, they migrate to the Saragasso Sea, on the west side of the Atlantic ocean, where they spawn and then die. The larvae then return back to Europe using the currents of the Gulf Stream[49]. Upon reaching the estuary, they reach a development stage, in which they are known as glass eels, due to their transparent appearance. The juvenile eels continue upstream where they then continue to mature and live.

It is known that European eels do not feed during their spawning migration[49]. As such, we hypothesise that the European eels do not interact significantly with the Guadalquivir estuary ecosystem. Due to the European eels only passing through the ecosystem, they are not included in the food-web. However, we can use the model we have to test for any causal interactions with other species. This is done using optimal values for the embedding dimension and lag.

The results of the CCM are shown in figure 5.10. In general, we can see that there are more green squares in figure 5.10b than figure 5.10a, suggesting that European eels interact more with the environment as juveniles than as adults. We can see that for the higher threshold value of 0.3 for adults and juveniles, there is little coupling with other species detected. However, for threshold values of 0.1 and 0.2, there is some coupling detected with other species at some of the sites and tides. From our model evaluation, we found that a threshold value of 0.1 and a required majority of 5 green squares has a precision of 54.2%. These criteria are passed by *Engraulis encrasicolus*, *Liza saliens* and *Mesopodopsis slabberi* interacting with the juvenile glass eels. This suggests that our hypothesis may be false since European eels appear to interact with some species, especially as juveniles. The reduction in interactions for adult European eels may be explained by them being less likely to be preyed on as they increase in size.

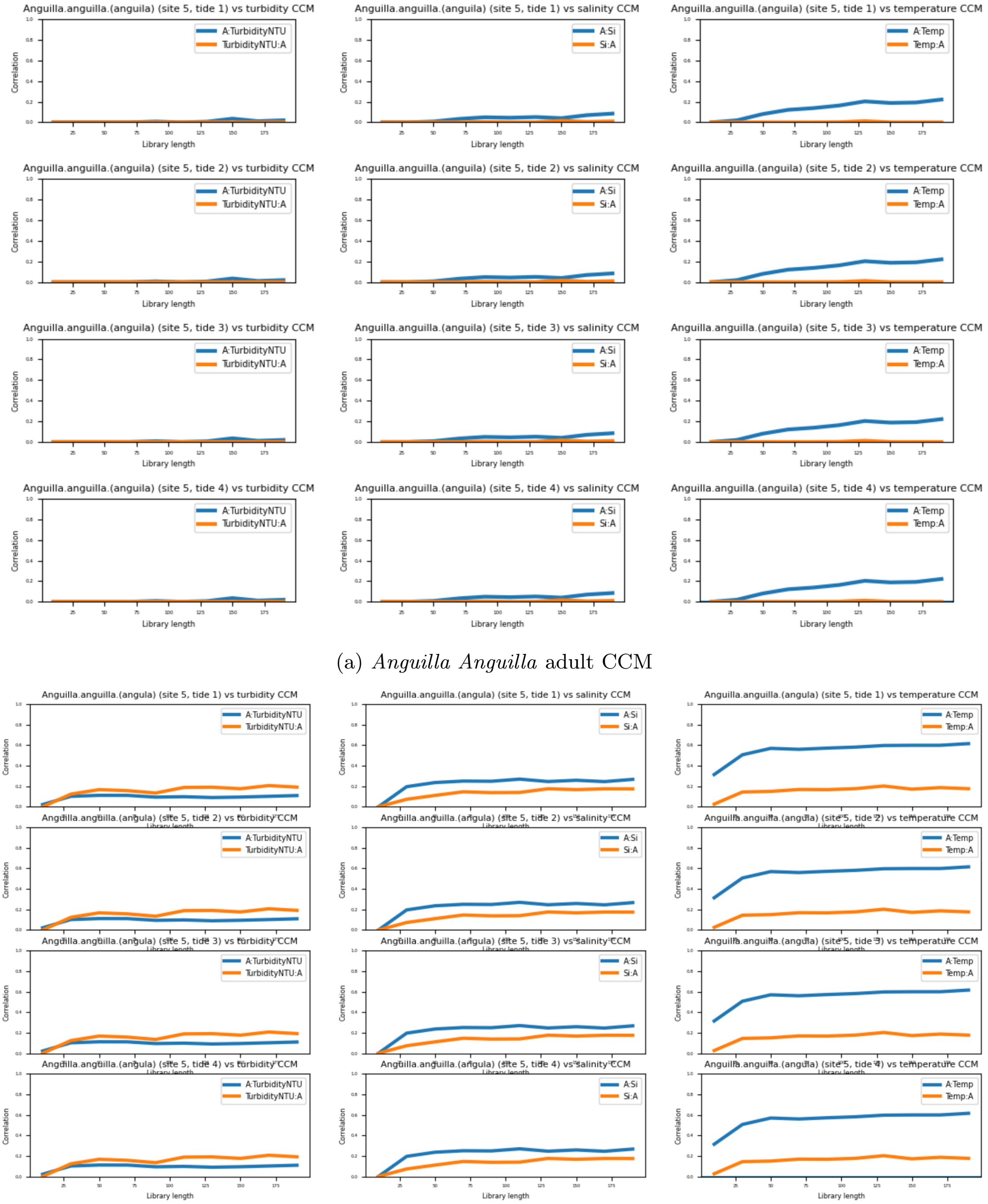


(a) Adult European eels

(b) Juvenile European eels

Figure 5.10: CCM between *Anguilla anguilla* (European eel) and top 10 species, using optimal values of E and τ. A green square indicates a correlation score above the threshold value in either direction.

We also test for causal interactions between the European eels and the environment, to examine if environmental factors may impact the abundance of the European eels. Since there is a higher abundance of European eels at site 5, than at site 3, we run CCM on the abundance of eels for each tide at site 5, against the environmental conditions.



(a) *Anguilla Anguilla* adult CCM

(b) *Anguilla Anguilla* juvenile CCM

Figure 5.11: CCM results of adult and juvenile *Anguilla anguilla* against environmental conditions

Figure 5.11a shows that adult European eels do not seem to be dynamically coupled with the turbidity or salinity of the estuary. However, the blue line converging to a non-zero value for the temperature plots at all tides suggests that the abundance of European eels is weakly driven by the temperature. The opposite CCM, represented by the orange line, remains at zero since the abundance of European eels is unlikely to drive the temperature. Figure 5.11b shows that turbidity and salinity have a weak influence on the abundance of glass eels, and temperature has a strong influence on their abundance. This may be due to them being more sensitive to environmental conditions during their migration as juveniles, or due to their migration being seasonal and the season strongly driving the environmental conditions. As shown in figure 2.2, all environmental factors change depending on the time of year. We notice that there is some weaker cross-estimate skill in the reverse direction CCM, with the results suggesting that the environmental conditions are driven by the European eel abundance. Since this is very unlikely, we can attribute this to synchronization and it is an example of what may give a false positive during the food-web predictions.

Chapter 6

Conclusion & Future Work

This project has presented an analysis of the Guadalquivir River estuary ecosystem and demonstrated an application of convergent cross mapping. We found that the estuary food-web exhibits small-world characteristics but is not scale-free. This ecosystem, predominantly occupied by juvenile species, was compared to other real-life food-webs and artificially generated networks. The CCM model has been shown to be somewhat effective at predicting the food-web: being able to produce a high level of recall, although with non-optimal precision. With such a model, similar large-scale food-webs can be semi-reliably predicted, reducing the need for expensive and time-consuming direct observation of species interactions.

The issues of the model can partially be attributed to the noise expected in most real-life datasets, measurement error, and the high degree of complexity in a dynamical system such as an ecosystem, leading to limited CCM convergence. The model could have potentially been improved further with more parameter fine-tuning, such as by testing larger ranges of embedding dimension and lag; but with the computing resources available the model already faced significant run-times, restricting the amount of parameters we could test. Despite the sample size having being shown to be sufficient, a larger sample size would have also increased CCM convergence, improving model performance.

As ecosystems have been shown to exhibit a variety of characteristics[1], CCM may have different performance on ecosystems with different levels of connectance and degree distribution. A food-web with a lower level connectance between species may have more defined coupling, making true interactions more distinct from false positives. This could be explored by comparing a CCM algorithm applied to multiple ecosystems, such as the ones listed in the study by Dunne et al.[1]. This project has also only focused on predicting a food-web using time-series information. Other methods exist for predicting species interactions, such as by inference from functional traits, phylogenies, and geography[50]. These methods could be considered and used in combination with CCM to predict the Guadalquivir River estuary food-web.

Future work could also focus on not only identifying interactions, but also quantifying the strength of the interactions. For example, the extent to which the CCM estimates converge can suggest the strength of interaction. By observing interaction strengths, species which are most likely to drive the dynamics of the entire ecosystem can be identified. This may be important, for example, if an invasive species is detected in an ecosystem and the extent of its threat to other species needs to be identified.

Another interesting study would to be identify the direct impact of dam discharges on the estuary ecosystem. There is an upstream dam on the Guadalquivir River that releases freshwater at varying rates. This freshwater has an impact on the turbidity and salinity of the water, which we have shown to drive the abundance of some species. With this information, policy on controlling dam discharges could be shaped to minimize negative environmental impacts.

Bibliography

[1] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. "Food-web structure and network theory: The role of connectance and size". In: *Proceedings of the National Academy of Sciences of the United States of America* (2002). ISSN: 00278424. DOI: 10.1073/pnas.192407699.

[2] Cesar Vilas. Private Communication. 2020.

[3] Mark Newman. *Networks: An Introduction*. 2010. ISBN: 9780191594175. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

[4] Paul Ehrlich. "Extinction: The causes and consequences of the disappearance of species". In: *Biological Conservation* (1983). ISSN: 00063207. DOI: 10.1016/0006-3207(83)90103-9.

[5] Jordi Bascompte and Daniel B. Stouffer. "The assembly and disassembly of ecological networks". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1524 (June 2009), pp. 1781–1787. ISSN: 14712970. DOI: 10.1098/rstb.2008.0226.

[6] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. "Network structure and biodiversity loss in food webs: robustness increases with connectance". In: *Ecology Letters* 5.4 (July 2002), pp. 558–567. ISSN: 1461-023X. DOI: 10.1046/j.1461-0248.2002.00354.x. URL: http://doi.wiley.com/10.1046/j.1461-0248.2002.00354.x.

[7] Jakob Runge. "Detecting and Quantifying Causality from Time Series of Complex Systems". PhD thesis. 2014, p. 255.

[8] George Sugihara et al. "Detecting Causality in Complex Ecosystems". In: *Science* 338.6106 (2012), pp. 496–500. ISSN: 0036-8075. DOI: 10.1126/science.1227079. eprint: https://science.sciencemag.org/content/338/6106/496.full.pdf. URL: https://science.sciencemag.org/content/338/6106/496.

[9] M. Carpintero et al. "Estimation of turbidity along the Guadalquivir estuary using Landsat TM and ETM+ images". In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology XV*. Vol. 8887. SPIE, Oct. 2013, 88870B. ISBN: 9780819497567. DOI: 10.1117/12.2029183.

[10] J. M. Vargas et al. "Seasonal and wind-induced variability of Sea Surface Temperature patterns in the Gulf of Cádiz". In: *Journal of Marine Systems* 38.3-4 (Jan. 2003), pp. 205–219. ISSN: 09247963. DOI: 10.1016/S0924-7963(02)00240-3.

[11] Jesús García-Lafuente et al. "Water mass circulation on the continental shelf of the Gulf of Cádiz". In: *Deep-Sea Research Part II: Topical Studies in Oceanography* 53.11-13 (June 2006), pp. 1182–1197. ISSN: 09670645. DOI: 10.1016/j.dsr2.2006.04.011.

[12] A. Reul et al. "Spatial distribution of phytoplankton <13 μm in the Gulf of Cádiz in relation to water masses and circulation pattern under westerly and easterly wind regimes". In: *Deep-Sea Research Part II: Topical Studies in Oceanography* 53.11-13 (June 2006), pp. 1294–1313. ISSN: 09670645. DOI: 10.1016/j.dsr2.2006.04.008.

[13] Javier Ruiz et al. "Meteorological and oceanographic factors influencing Engraulis encrasicolus early life stages and catches in the Gulf of Cádiz". In: *Deep-Sea Research Part II: Topical Studies in Oceanography* 53.11-13 (June 2006), pp. 1363–1376. ISSN: 09670645. DOI: 10.1016/j.dsr2.2006.04.007.

[14] Eugene P. Odum. "The strategy of ecosystem development". In: *Science* 164.3877 (1969), pp. 262–270. ISSN: 00368075. DOI: 10.1126/science.164.3877.262.

[15] Dafeng Hui. *Food Web: Concept and Applications*. 2012. URL: https://www.nature.com/scitable/knowledge/library/food-web-concept-and-applications-84077181/.

[16] T. M. Smith and Robert Leo Smith. *Elements of Ecology*. Boston, Massachusetts: Pearson, 2015, p. 706.

[17] Jonathan B. Shurin, Daniel S. Gruner, and Helmut Hillebrand. *Review All wet or dried up? Real differences between aquatic and terrestrial food webs*. Jan. 2006. DOI: 10.1098/rspb.2005.3377.

[18] Just Cebrian. “Role of first-order consumers in ecosystem carbon flow”. In: *Ecology Letters* 7 (2004), pp. 232–240. DOI: 10.1111/j.1461-0248.2004.00574.x.

[19] Robert T. Paine. *Food Web Complexity and Species Diversity*. Tech. rep. 910. 1966, pp. 65–75. DOI: 10.1086/282400. URL: https://about.jstor.org/terms.

[20] Russell J. Schmitt. *Indirect Interactions Between Prey : Apparent Competition , Predator Aggregation , and Habitat Segregation*. Tech. rep. 6. 1987, pp. 1887–1897.

[21] Nelson G Hairston, Frederick E Smith, and Lawrence B Slobodkin. *Community Structure, Population Control, and Competition*. Tech. rep. 879. 1960, pp. 421–425.

[22] D. Irvin Rasmussen. *Biotic Communities of Kaibab Plateau, Arizona*. Tech. rep. 3. 1941, pp. 229–275. DOI: 10.2307/1943204.

[23] Joel E. Cohen, Frédéric Briand, and Charles M. Newman. *Community Food Webs*. Vol. 20. Biomathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990. ISBN: 978-3-642-83786-9. DOI: 10.1007/978-3-642-83784-5. URL: http://link.springer.com/10.1007/978-3-642-83784-5.

[24] Eugene P. Odum. “Energy flow in ecosystems: A historical review”. In: *Integrative and Comparative Biology* 8.1 (1968), pp. 11–18. ISSN: 15407063. DOI: 10.1093/icb/8.1.11.

[25] Stanley Milgram. “The Small World Problem”. In: *Psychology Today* 1.1 (1967), pp. 61–67.

[26] Duncan J. Watts and Steven H. Strogatz. “Collective dynamics of ’small-world’ networks”. In: *The Structure and Dynamics of Networks*. Vol. 9781400841. Princeton University Press, Oct. 1998, pp. 301–303. ISBN: 9781400841356. DOI: 10.1038/30918.

[27] Albert László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *Science* 286.5439 (1999), pp. 509–512. ISSN: 00368075. DOI: 10.1126/science.286.5439.509.

[28] JOSE M. MONTOYA and RICARD V. SOLÉ. “Small World Patterns in Food Webs”. In: *Journal of Theoretical Biology* 214.3 (2002), pp. 405–412. ISSN: 0022-5193. DOI: https://doi.org/10.1006/jtbi.2001.2460. URL: http://www.sciencedirect.com/science/article/pii/S0022519301924609.

[29] Neville Davies and C. Chatfield. *The Analysis of Time Series: An Introduction*. 6th ed. Vol. 74. 468. CRC Press, Mar. 1990, p. 194. ISBN: 9780203491683. DOI: 10.2307/3619403.

[30] John Gubner. *Probability and Random Process for Electrical and Computer Engineers*. Cambridge University Press, 2006, p. 392. ISBN: 9780521864701. URL: www.cambridge.org/9780521864701..

[31] Robert W. Floyd. “Algorithm 97: Shortest Path”. In: *Commun. ACM* 5.6 (June 1962), p. 345. ISSN: 0001-0782. DOI: 10.1145/367766.368168. URL: https://doi.org/10.1145/367766.368168.

[32] Thomas H. Cormen et al. *Introduction to Algorithms, Third Edition*. 3rd. The MIT Press, 2009. ISBN: 0262033844.

[33] Gert Sabidussi. “The centrality index of a graph”. In: *Psychometrika* 31.4 (1966), pp. 581–603. URL: https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:4:p:581-603.

[34] Giorgio Fagiolo. “Clustering in complex directed networks”. In: *Physical Review E* 76.2 (Aug. 2007). ISSN: 1550-2376. DOI: 10.1103/physreve.76.026107. URL: http://dx.doi.org/10.1103/PhysRevE.76.026107.

[35] C. W. J. Granger. “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”. In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

[36] Kevin Mccann, A.G. Hastings, and Gary Huxel. “Weak Trophic Interactions and the Balance of Nature”. In: *Nature* 395 (Oct. 1998), pp. 794–798. DOI: 10.1038/27427.

[37] Anatole Katok and Boris Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1995. DOI: 10.1017/CBO9780511809187.

[38] Milton Persson. "The Whitney embedding theorem". PhD thesis. 2014. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-91352.

[39] Floris Takens. "Detecting strange attractors in turbulence". In: *Dynamical Systems and Turbulence, Warwick 1980*. Ed. by David Rand and Lai-Sang Young. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, pp. 366–381. ISBN: 978-3-540-38945-3.

[40] Han Lun Yap. *Takens' Embedding Theorem*. 2011. URL: http://cnx.org/contents/939eff33-c4f0-43e4-ae2b-ca6d5b58e267@2.

[41] Edward N. Lorenz. "Deterministic Nonperiodic Flow". In: *Journal of the Atmospheric Sciences* 20.2 (Mar. 1963), pp. 130–141. ISSN: 0022-4928. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. eprint: https://journals.ametsoc.org/jas/article-pdf/20/2/130/3406061/1520-0469(1963)020_0130_dnf_2_0_co_2.pdf. URL: https://doi.org/10.1175/1520-0469(1963)020%3C0130:DNF%3E2.0.CO;2.

[42] Ethan R. Deyle and George Sugihara. "Generalized Theorems for Nonlinear State Space Reconstruction". In: *PLOS ONE* 6.3 (Mar. 2011), pp. 1–8. DOI: 10.1371/journal.pone.0018295. URL: https://doi.org/10.1371/journal.pone.0018295.

[43] SugiharaLab. *pyEDM*. https://github.com/SugiharaLab/pyEDM. 2020.

[44] nickc1. *skccm*. https://github.com/nickc1/skccm. 2018.

[45] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. 2nd ed. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511755798.

[46] David Chelidze. *Delay Coordinate Embedding*. URL: https://personal.egr.uri.edu/chelidz/documents/mce567_Chapter_7.pdf (visited on 06/14/2020).

[47] Masayuki Ushio et al. "Fluctuating interaction network and time-varying stability of a natural fish community". In: *Nature* 554 (Feb. 2018). DOI: 10.1038/nature25504.

[48] Kim Aarestrup et al. "Oceanic Spawning Migration of the European Eel (Anguilla anguilla)". In: *Science (New York, N.Y.)* 325 (Sept. 2009), p. 1660. DOI: 10.1126/science.1178120.

[49] Fao. *Cultured Aquatic Species Information Programme. Anguilla anguilla*. URL: http://www.fao.org/fishery/culturedspecies/Anguilla_anguilla/en#tcN90078 (visited on 06/14/2020).

[50] Ignacio Morales-Castilla et al. "Inferring biotic interactions from proxies". In: *Trends in Ecology Evolution* 30.6 (2015), pp. 347–356. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2015.03.014. URL: http://www.sciencedirect.com/science/article/pii/S0169534715000774.