# Forecasting Water Pollutants

António Gonçalo Fontes Pinheiro
Mestrado em Ciência de Computadores
Departamento de Ciência de Computadores
2018

**Orientador**
Rita Paula Almeida Ribeiro, Professora Auxiliar,
Faculdade de Ciências da Universidade do Porto

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

# Abstract

Nowadays, a huge number of emerging pollutants coming from different sources, such as industrial activities and man-made products, enter our rivers from the Waste Water Treatment Plants (WTTPs) effluent. The reason for this is that these new type of pollutants are not currently removed by the WWTPs. As such, the probability of them entering into the environment is dangerously high, causing a great impact in human health and in the aquatic organisms. Even though, their identification is currently made in some rivers and WTTPs, the knowledge about how to effectively identify and remove them is still too small. Thus, it is crucial to spread the information about emerging pollutants and increase the awareness of their impact in the environment. Therefore, it is essential that new methods are developed to automate the detection and prediction of concentration values of these pollutants so that some preventive and corrective actions can be taken.

The main goal of this thesis is to study different forecasting methods to anticipate the concentration values of emerging pollutants. We developed and tested several prediction models in two different case studies.

In the first case study, we used data concerning pharmaceuticals concentrations found at the source and WTTPs of the Lis river, Portugal. For this case study, we tested some of the most common forecasting methods and explored the incorporation of the hierarchical structure of the pharmaceuticals. Despite being an interesting and highly relevant problem, the small amount of available data has shown to be major limitation to this study.

In the second case study, we used publicly available data regarding other pollutants found in rivers in the United States of America. For this case study we explored the use of standard forecasting methods, regression methods on embedded time series and the incorporation of the spatial relationship of the rivers in the prediction method. Results have shown that, as expected, the error estimates obtained were much lower than in the first case study while using the same prediction. The use of embedded time series allowed us to obtain better results and take advantage of the spatial components in the Lithium data set, where it was possible to prove its benefit in the prediction task.

Keywords: Emerging Pollutants, Data Mining; Time Series

# Resumo

Hoje em dia, um grande número de poluentes emergentes provenientes de diferentes origens, tais como, atividades industriais e produtos fabricados pelo homem, entram para os nossos rios através dos efluentes das Estações de Tratamento de Águas Residuais (ETARs). Isto acontece porque este novo tipo de poluentes não são atualmente monitorizados pelas ETARs. Por isso, a probabilidade deles entrarem no nosso ambiente é perigosamente alta, causando um grande impacto na saúde humana e nos ecossistemas aquáticos. Apesar da identificação dos poluentes emergentes já estar a decorrer nos rios e ETARS, o conhecimento sobre como os identificar e remover com eficácia é ainda muito escasso. Portanto, é essencial que, o mais rapidamente possível, melhoremos os métodos para a sua identificação e deteção para que seja possível dar resposta a este novo problema. Portanto, é crucial divulgar a informação sobre os poluentes emergentes e promover a consciencialização do público sobre o seu impacto no meio ambiente. É essencial desenvolver novos métodos para automatizar a deteção e previsão da concentração destes poluentes, de forma a que ações preventivas e corretivas possam ser tomadas.

O objetivo principal desta tese é estudar diferentes métodos de previsão para antecipar os valores das concentrações destes poluentes emergentes. Desenvolvemos e testamos diferentes métodos de previsão em dois casos de estudo diferentes.

No primeiro caso de estudo, nós usamos dados relativos a concentração de fármacos encontrados na nascente e nas ETARs do rio Lis em Portugal. Para este caso de estudo, testamos alguns dos mais comuns métodos de previsão e exploramos o uso de uma estrutura hierárquica dos fármacos. Apesar de ser um problema interessante e bastante relevante, a quantidade pequena de dados disponíveis mostrou ser uma das grandes limitações a este estudo.

No segundo caso de estudo, usamos dados disponíveis publicamente sobre outros poluentes encontrados nos rios dos Estados Unidos da América. Para este caso de estudo, exploramos o uso dos métodos comuns de previsão, métodos de regressão em séries embebidas e aproveitamos a relação espacial dos rios nos métodos de previsão. Os resultados demonstraram que, como esperado, as estimativas de erro obtidas foram bastante mais pequenos que no primeiro caso de estudo para os mesmos métodos de previsão. O uso de séries embebidas permitiu-nos obter melhores resultados e aproveitar os componentes espaciais dos dados do Lithium, onde foi possível provar os benefícios do seu uso na tarefa de previsão.

Palavras-chave: Poluentes Emergentes, Data Mining, Séries Temporais

# Agradecimentos

Um agradecimento especial:

- À Professora Rita Ribeiro pela orientação prestada, pela sua disponibilidade, sugestões e apoio durante os períodos mais complicados da tese. Acima de tudo, obrigado pela motivação dada ao longo da tese e pela simpatia demonstrada.

- À Professora Cristina Delerue-Matos e à Dra. Paula Paíga por toda a ajuda e disponibilidade ao longo do projeto.

- À minha família por sempre acreditarem em mim e me apoiarem durante a minha vida pessoal e académica.

- À Inês, por todo o carinho, paciência e incentivo em todos os momentos da minha vida.

- À todas as pessoas que marcaram o meu percurso académico e estiveram sempre presentes, não só nos momentos mais divertidos e de convívio, mas também nos momentos mais complicados.

- Este trabalho foi apoiado pela FCT no âmbito do projeto FOTOCATGRAF (UTAP-ICDT/CTM-NAN/0025/2014).

**Dedicada ao meus Pais e à Inês.**

# Contents

# List of Tables

# List of Figures

# Acronyms

| | |
|---|---|
| **ACF** | Autocorrelation Function |
| **ADE** | Arbitrated Dynamic Ensemble |
| **AIC** | Akaike Information Criterion |
| **ANN** | Artificial Neural Network |
| **ARIMA** | Auto-Regression Integrated Moving Average |
| **ETS** | Error, Trend, Seasonality Method |
| **HTS** | Hierachical Time Series |
| **KDD** | Knowledge Discovery and Data Mining |
| **MAE** | Mean Absolute Error |
| **MAPE** | Mean Absolute Percentage Error |
| **MDL** | Method Detection Limit |
| **MQL** | Method Quantitation Limit |
| **SES** | Simple Exponential Smoothing |
| **STP** | Sewage Treatment Plants |
| **SVM** | Support Vector Machine |
| **TBATS** | Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality |
| **USGS** | United States Geological Survey |
| **WWTP** | Wastewater Treatment Plant |
| **WWTPs** | Wastewater Treatment Plants |

# Chapter 1

# Introduction

Nowadays, more and more, pollutants enter the aquatic ecosystems from the Wastewater Treatment Plant (WWTP) effluents due to the population growth combined with the industrial activities. The main topic of this thesis focus on the need of improving and creating new detection methods of such pollutants in our rivers. In this chapter we will describe our motivation, main goals and how we decided to organize this thesis.

## 1.1 Motivation

Water is the most crucial resource for human survival. Without it, we would die in a matter of days (Edition, 2011). Besides that, it is an essential element in many industries like for example the agricultural industry. The need to keep pace with new challenges that appear everyday caused by the industrial expansion and human population growth led to a big rise in industrial pollution that sadly has been having a significant impact in our water quality (Flores Alsina et al., 2008). For this reason, there is a growing need to build and ensure the effectiveness of Wastewater Treatment Plants (WWTPs).

WWTPs promote the reuse of water previously used both by industry and people, while ensuring the public health by eliminating the waste and nearly all pollutants before the effluent is released (Spellman, 2013).

### 1.1.1 Emerging Pollutants

Over the last decade, with the recent advances in analytical methodologies, we have been able to detect low concentrations of xenobiotic man-made chemicals which were produced and are now found in our drinking water and in aquatic ecosystems (Thomaidis et al., 2012). This low levels contaminants are called "emerging pollutants" and they include a wide range of chemicals (such as pharmaceuticals, pesticides, metabolites, among others) which are used daily around the world (Gavrilescu et al., 2015).

Presently, the methods used in water purification by WWTPs are not effective in the removal of emerging pollutants. Although it is easy to describe what they are, when it comes to identifying them we can no longer say the same. These pollutants are known to be persistent in the environment and complex in their forms and actions, making the identification difficult. As result of being so recent, there is a lack of awareness among the public, industries and governments, thus making it one of the main objectives of the United Nations: to report their risks and dangers, as well as suggest ways of how to combat them.

Taking into account that there is few data on these new pollutants, it is crucial to increase the collection and sharing of new data so that new removal techniques can be studied. The design of such techniques is not straightforward. These emerging pollutants are not regulated, either by national or international laws and since they comprise a wide range of different contaminants, it is difficult to know which ones to target. Moreover, without the mandatory technical capabilities, we can not handle the unique requirements they demand. They are present in low concentrations in our drinking water but, even though it is not currently a threat to our safety, we do not have sufficient information to be sure of what will be their impact in the near future (Geissen et al., 2015).

### 1.1.2 Pharmaceuticals and other contaminants in WWTPs

WWTPs and Sewage Treatment Plants (STP) were not designed to attack and efficiently remove pharmaceuticals and other type of contaminants (e.g. personal care products) from the aquatic environment. This absence of action has obvious repercussions, such as a growth in the detection rate of these pollutants in rivers as well as in the effluents of treatment plants.

Specially on rivers dominated by effluent flows, the lack of treatment has an impact in the water quality characteristics that, to a certain degree, differ from a river in normal conditions. It is important to notice that in some rivers, in the dry periods, the effluents can have a strong influence in the aquatic environment, since the percentage of water originated from waste waters is much higher (Roberts et al., 2016).

It is known that aquatic organisms and environments can handle sudden increases in the levels of pollution during a short time, for example caused by a rainfall event, but continuous and permanent exposure to the high levels of contaminants can produce severe changes in organism's processes such as reducing fertility (Brooks et al., 2006).

In this context, it is important that new methods are developed to detect and forecast these emergent pollutants.

## 1.2 Goals

This thesis is integrated in the context of the FOTOCATGRAF research project, which aims at developing a new technologies for the removal of emerging pollutants from wastewater to ensure a safer and sustainable water supply. This problem is one of the grand global challenges of the 21st century. The growth of population and industry makes it crucial to take action. To address this problem, the goal of the project is to produce a new generation of high-performance graphene-based photocatalysts to make the removal of emerging pollutants more efficient and accurate.

One of the tasks to accomplish such goal is to develop a data mining system to monitor and predict pollutants patterns using the data collected from WWTPs of the center region of Portugal. The objective is to forecast concentrations of certain pollutants so that, in the future, preventive actions can be taken for their removal before it is too late. The scope of this thesis focuses on this task.

## 1.3 Organization

This thesis is organized in five chapters, briefly, described next.

**Introduction -** The current chapter provides an overview about the motivation for this study and the main established goals.

**Background -** This chapter introduces background knowledge on time series, their characterization and standard forecasting methods; it also introduces the tools used in this study.

**Case Study 1: Forecasting Pharmaceuticals Pollutants in Water -** This chapter focuses on the problem that motivated this study; a detailed description about the target pharmaceuticals and their conditions on Lis river, Portugal, is given, some forecasting methods are experimented and the limitations found for this case study are discussed.

**Case Study 2: Forecasting other Pollutants in Water -** This chapter presents the forecasting problem of three different contaminants found in rivers in United States of America; it describes the implementation of forecasting techniques that were not possible to use on the first case study; and, a discussion of the results obtained from the experiments is also given.

**Conclusions -** Finally, this chapter presents the main conclusions that we can draw from this study, limitations that we had and what can we improve in future work.

# Chapter 2

# Background

This chapter introduces the main basic concepts necessary for the comprehension of this study. We start by explaining what is data mining. Then, we introduce the definition of time series, the several components that build it and the transformations that can be applied in order to help a better forecast. Next, the forecasting methods applied in this work are briefly described. In the last section, we mention all the tools that were used along this study, both for the pre-processing and the forecasting tasks.

## 2.1 Data Mining

Now, more than ever, our society is at the mercy of information and the innumerable ways to get it. This results from the massive growth that technology has been having over the last decades and the constant need to store information. What was once before registered in paper and then stored in warehouses is now in clouds or huge databases. It is important to not waste such information that usually is stored in raw form and in an unstructured way, such as facts of transactions or a format that we cannot process right away without previous pre-processing. Information can be a very useful and an important tool for a company but, if no action is taken, it is useless and can cause the company to fall behind the competition (Witten et al., 2016).

According to Hand (2007), data mining is the analysis of observational data stored in databases, to find unknown relationships and summarize the data in a new way that can be understandable and valuable to the respective owner. Usually it is applied to large data sets since if we had only a low number of records to analyze there would be no need to create some type of automation and it would be a rather simple task to a statistician.

There are several real-world applications where data mining techniques can have crucial role for handling continuously generated data. Larose and Larose (2014) give an example of a supermarket. Each time a customer goes to the cashier to pay for the products, each one of them is scanned and the database is consulted to check for the price and possible promotions. But that is not the only interaction made with the database. Every time you buy a product, they save the record of the sale, the time of the day you went to the store, how many units of the product you bought, what products you also bought at the same time and, of course, the information associated with you that can be later used to calculate the mean of age of people buying a certain type of product. Every interaction generates a new row in the data base, and if we multiply that by the number of items in your shopping cart, and then multiply the result by the number of interactions happening at the same time by the rest of the cashiers, we can see how big the amount of data generated is in such a tiny amount of time.

Handling and extracting valuable information from such amount of data is not easy and, for that purpose, a technique named Knowledge Discovery and Data Mining (KDD) is used. For some people KDD is a synonym of data mining, while for others data mining is one of the most important steps of the KDD process. According to Han et al. (2011), we can divide the KDD process in 7 steps as described next.

1. Data integration – Data from multiple data sources is combined into one single data set.
2. Data cleaning - Noise or irregularities are removed from the data and missing values are handled.
3. Data selection – In this step, the relevant data for the KDD task is selected and dimensionality reduction methods are applied in order to find invariant representations of the data and remove redundant variables (Fayyad et al., 1996).
4. Data transformation – Operations, such as aggregation and summarization, are applied to obtain a data set in a specific format.
5. Data mining – In this step methods are applied to the transformed data in order to extract patterns in a specific representation form like classification rules, clusters or regression.
6. Pattern evaluation – Visualization of the previous extracted patterns by the data mining method. With the new information obtained from the interpretation and evaluation of the extracted patterns more iterations might be necessary and, in that case, we return to the previous steps.
7. Knowledge presentation – Based on the accomplished knowledge report it to the interested parties and decide what is the next action given the initial goals of the KDD process.

According to Turban et al. (2008), the data mining tasks performed on the $5^{th}$ step of the KDD process can be divided into two main tasks.

1. Predictive Tasks - Technique used to predict unknown or future values by trying to find certain patterns that help us to understand the behavior of a certain variable.
   (a) Classification - Task of predicting a categorical variable, like a specie of a plant by her characteristics;

   (b) Regression - Task of predicting a numeric variable, for example, how many customers will buy a specific product.

2. Descriptive Tasks - Technique used to derive patterns (Clusters, Correlations, and others) that show the relationships behind the data.

   (a) Associations – Technique used in data mining to discover relations between different variables in large databases, relations that without this method we would not be able to see, at least, without wasting too much time analyzing huge numbers of records.

   (b) Clustering – Divides objects into separated groups in which every member shares similar characteristics. Since different algorithms will give different clusters it is essential to interpret if the set of clusters, that were generated from our data set, is relevant and logical and, only after that, we can use the newly created clusters to analyze new data.

Data mining is a popular used tool in the business world and can be used in variety contexts as has already been demonstrated in this chapter. Many businesses use data mining to solve problems that are affecting them, whether they are logistical or financial, while others just want to gain advantage over competition by trying to be one step ahead Turban et al. (2008).

Some examples of data mining application domains include:

- customer relationship management, where the previous transactions of a customer are used in order to try to comprehend his needs and build an one-on-one relationship;
- banking, where credits and clients history is used to predict what would be the best loan for a client within his possibilities;
- retailing and logistics, where history of sales is taken into account to predict necessary inventory levels in each store;
- computer software, where programs such as anti-virus use data mining to identify possibly infected software;
- health care, where data mining methods are used to, for example, predict potential overload of patients during critical times like winter;
- medicine, where one of the most common tasks uses data mining to understand relationships between symptoms and diseases to help predict and treat future illness;
- earth sciences, where the goal is to use data mining methods to support decision making for environmental management, such as to control pollution levels or the excessive growth of harmful algae.

## 2.2 Time Series

For some application domains, the data set is considered to be composed by observations that are independent from each other. Still, that is not the case for many real-world domains where observations actually are related to each other temporally, spatially or by another type of relationship. For example, observations generated by a temperature sensor, or by electrocardiogram have a clear temporal dependence. This dependence needs to be explicitly accounted for in the data mining process. The prediction model must be able to model the temporal dependencies between the values in order to successfully forecast the next values.

A time series is a sequence of values measured successively during a specific time interval and mathematically it can be defined by the following equation:

$$Y = (y_1, y_2, \ldots, y_n), y_t \in \mathbb{R} \tag{2.1}$$

where $t$ represents the time at which the value $y_t$ was observed.

If the measured values are equally separated in time (e.g. every hour or month), we have a regular time series. If the time interval between each observation is not always equal (for example, if it depends on sensors for certain external conditions, such as the log of a machine), we have an irregular time series.

Analyzing a time series can have many applications like stock market analysis, inventory management, sales forecasting, ecological modelling and so on.

### 2.2.1 Components of a Time Series

Decomposition is an important tool used for time series analysis that allows to separate a series into several components that help us to identify patterns, relationships and behaviours that can be invisible to the naked eye. This distinct components, with different properties, can help not only on the analysis task but also with the forecast task since we can get a better understanding on how the behaviour is influenced by trend, seasonality or other time series components. Overall, we can identify four main different components in a time series, as described next.

**Trend component -** It happens when there is a long-term increase or decrease in the time series values. This is typically observed in time series that are related to population growth or price inflation;

**Seasonal component -** This pattern occurs if the time series is influenced by the calendar. For example, the increase of sales on a toy shop in December because of Christmas time, or the growth of tourists in the summer;

**Cyclic component -** This happens when the time series shows cyclical variations. This component is non-seasonal and varies in a recognizable cycle, occasionally showing an oscillation without a fixed period but predictable to a certain point. One example of this can be economic data affected by expansion or recession phases of economic systems.

**Irregular component -** This is what remains after we remove the trend and seasonal components from the time series. The irregular component, also called residuals, may or may not be random and it is the result of several fluctuation that, in case of a very irregular series, can dominate movements and hide the trend and seasonality components, making it harder to identify them.

Figure 2.1 shows the decomposition of an original time series into three components, where the last component includes the irregular and the cyclic component.

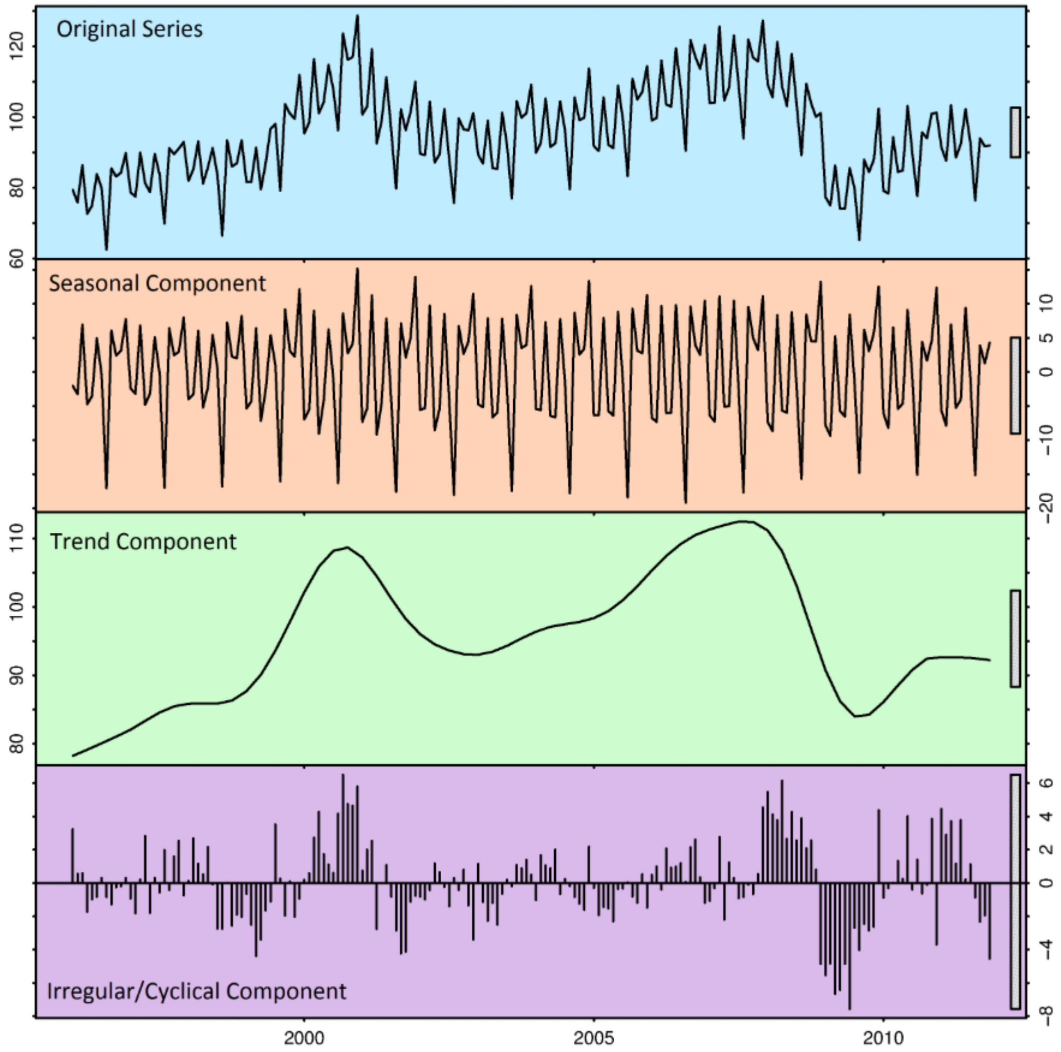


Figure 2.1: Decomposition of a time series (Libesa, 2014)

We can define the time series as a function containing the components mentioned above plus the error. Depending on whether we assume an additive or multiplicative model, the time series can be defined by Equation 2.2 or Equation 2.3, respectively.

$$y_t = S_t + T_t + E_t \tag{2.2}$$
$$y_t = S_t \times T_t \times E_t \tag{2.3}$$

where the $y_t$ is the value at the time $t$, $S_t$ is the seasonal component at time $t$, $T_t$ is the aggregation of the trend and cycle components at time $t$ and $E_t$ is the error at time $t$.

The additive model is mostly used when the seasonal amplitude is constant over time, meaning that, for example, the difference between two months is pretty much the same every year. On the other end, the multiplicative model gives more importance to the percentage changes instead of differences, like for example, looking if a specific month when compared to another is proportionally higher or lower constantly every year.

### 2.2.2 Transformations

A time series is defined by a sequence of values that are in chronological order. In stationary time series it is assumed that each value of the variable $Y$ at time $t$, $y_t$, is independent from each other. Nevertheless, this property is not usually observed. Any time series that shows trend and/or seasonality cannot be denominated as stationary given that, if one of those two properties is present, the values at specific periods of time are affected by them. One way of identifying a stationary series is when any period of time you choose, the values will look the same (NIST).

The time series shown in Figure 2.2, that represents the number of monthly totals of international airline passengers from 1949 to 1960, is an example of a non-stationary series. We can conclude that just by observing the change of the mean and the variance over time.

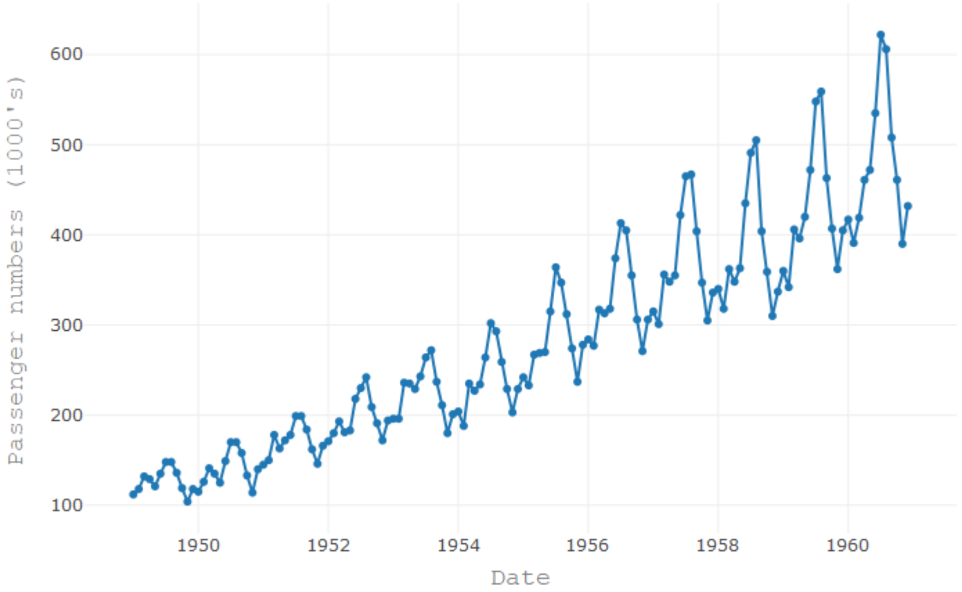


Figure 2.2: Air Passengers time series from 1949 to 1961

#### 2.2.2.1 Power transformations

Power transformations is a set of power functions applied to data in order to, for example, stabilize variance, give the data a normal distribution. Some common examples of such power transformations include:

**logarithmic scale:** by applying the logarithmic function to every value, a time series with a exponential distribution can become linear this transformation can be used in time series with an exponential distribution to make it linear.

**square root:** by applying the square root function to every value, a time series with a quadratic growth can become linear.

In addition to these two power functions there are other functions with the same goal of normalizing the data.

#### 2.2.2.2 Imputation Methods

Time series with missing data can be a big barrier when trying to create a forecasting model, since some methods require complete data and the lack of data can cause problems when trying to understand the trend and seasonality. Specially in non-automated sampling methods that need human intervention, the possibility of missing data is much greater. Therefore, it is frequently necessary to replace all the missing values on the data set by using an imputation method. This method should take into account the characteristics of the time series, in order to generate appropriated and rational values (Elissavet, 2017).

A common approach to this problem is the replacement by weighted moving average. In this method the missing values get replaced by moving average values. The mean used is taken from an equal number of observations on either side of the missing value. So, if the missing value is at timestamp position $t$ and the window size is 2, we will take the observations of timestamps $t-2$, $t-1$, $t+1$ and $t+2$ to calculate the mean. If the time series has consecutive missing values the window size is increased until we have at least two non-missing values inside the window (Moritz and Bartz-Beielstein, 2017).

Other imputation methods include the Last Observation Carried Forward (LOCF), where the missing value is replaced by the most recent non-missing value, linear interpolation, or not time series specicif methods, like for example, using the mean or linear regression.

#### 2.2.2.3 Differencing

A common approach to convert a time series to a stationary form is differencing. When we apply differencing to a series we obtain a new series that is composed by the consecutive changes between the different values in the time series, i.e.

$$y'_t = y_t - y_{t-1} \tag{2.4}$$

where $y'_t$ is the new value at the time $t$ obtained from the original values of the series at times $t$ and $t - 1$.

One thing to notice is that the first observation will always be lost since it is impossible to difference from the past value for the first value.

In Figure 2.3 we can see what changes after the differencing on the non-stationary series shown in Figure 2.2.

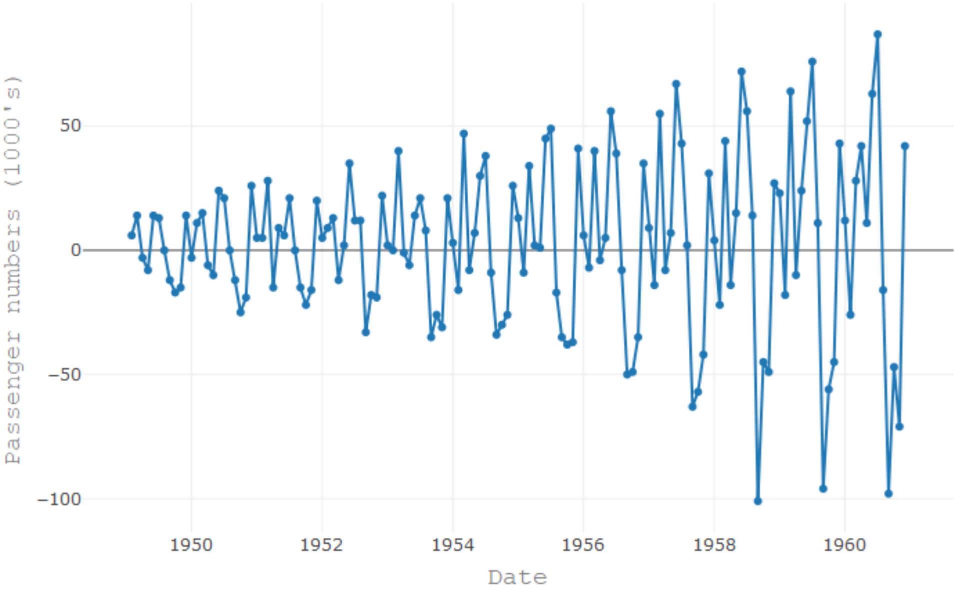


Figure 2.3: Differencing applied to the air passengers time series from 1949 to 1961

If, after the first differencing, the data is still non-stationary we can try to apply in the original series a second-order differencing that consists on using the change between the past two differences, causing us to loose the first two values. This second-order differencing can be written as

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned} \tag{2.5}$$

### 2.2.3 Autocorrelation Function

A good approach to determine if a series is stationary is using an Autocorrelation Function (ACF). The ACF measures the correlation between observations of a time series that are separated by $k$

time units ($y_i$ and $y_{i+k}$) and can be obtained by:

$$r_k = \frac{\sum_{i=1}^{N-k}(y_i - \overline{Y})(y_{i+k} - \overline{Y})}{\sum_{i=1}^{N}(y_i - \overline{Y})^2} \tag{2.6}$$

where $N$ is the size of the time series and $\overline{Y}$ represents the mean value of the time series.

Through the ACF it is also possible to confirm that differencing made the series stationary and, if not, showing that there is a need for a second-order differencing. For example, in Figure 2.4, we have the ACF plot of a non-stationary series. As it is possible to observe, the ACF is slowly decreasing while staying higher than the significance range, represented by the dashed blue lines. After the differencing the same series, if we use ACF again We obtain the graph shown in Figure 2.5. This time the ACF value decreases very fast on the first lag values and stays between the significant range. This behavior is characteristic of a stationary series because, independently of the lag, you are looking the correlation should be zero or very close to it.

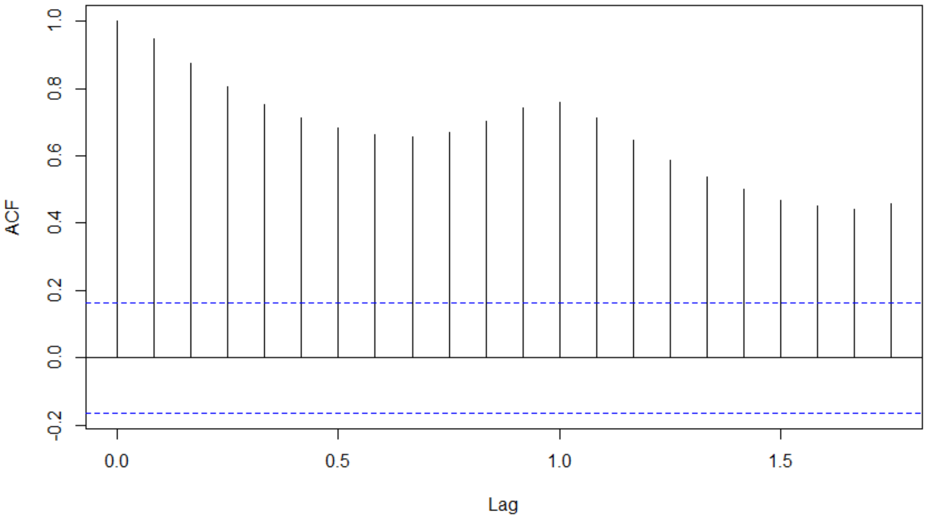


Figure 2.4: ACF of the AirPassengers data set

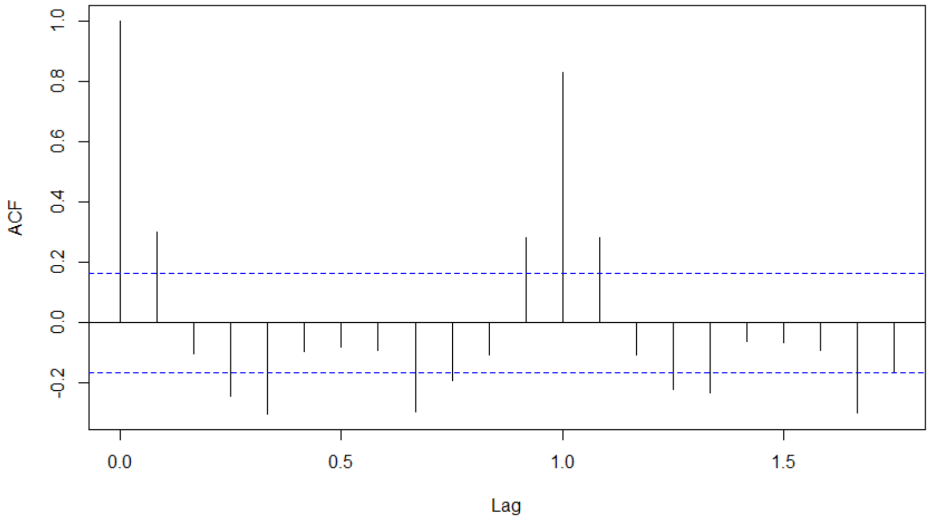


Figure 2.5: ACF of the AirPassengers data set after Differencing

## 2.3 Time Series Forecasting

One of the most common applications of time series analysis is forecasting, i.e. use the history of previous values to predict future values. The prediction of future trends is relevant in many real-word domains, such as weather forecast, economic indicators, retail sales and stock market.

If the time series is stationary, then estimated statistical parameters such as mean, variance or statistical correlation of the series will not change significantly over different time windows. Thus, in this case, such parameters are good indicators of the future behavior of the series. Nevertheless, this is not true if the time series is non-stationary. This is why is often advantageous to convert non-stationary series to stationary ones before forecasting. Some of the following forecasting models assume a stationary time series, whereas others do not.

### 2.3.1 Linear Models

Linear prediction models assume a linear relationship between parameters of the variables. That is, a change on one of the parameters is independent of another. These are simple and easy to understand models.

#### 2.3.1.1 Random Walk

Starting with the simplest forecast model, we have the random walk model. This method says that what happened before will happen again, and it uses that to forecast future values. It is a very straightforward method, very cheap to compute and sometimes achieves better results when compared to more complex forecast models.

The good thing about random walk approach is that it can be used as a baseline to test more complicated models since, if the chosen model does not outperform the random walk model, then it should not even be considered.

If the series is stationary, this method will use the last measured value as a forecast for the next values, but if somehow the series shows seasonality variations, then the forecast will depend on the last season on the series. This is the case when it is assumed that the value of tourists on a city during the summer is equal to the number of tourists during the previous summer or that the number of sales in a shop on Mondays is equal to the amount of sales in last week Monday and so on.

#### 2.3.1.2 Moving Average

The random walk approach lacks in the fact that it only considers a lag of one period, making the capture of possible trends harder. This can be solved by using lag values higher than one, and for that we have the moving average. In this method, to forecast the next values, we use the average of the last $n$ observations before the period of forecast, where $n$ is defined by the user and each observation receives equal weight in the forecast. As so, the moving average forecast is defined the following equation:

$$\hat{y}_t = \frac{\sum_{i=1}^{n} y_{t-i}}{n} \tag{2.7}$$

where $\hat{y}_t$ is the forecast for the time $t$ and $y_i$ is the actual value in time $i$. To achieve a bigger smoothing effect we can increase the number of forecasts in the moving average.

#### 2.3.1.3 Weighted Moving Average

The weighted moving average is very similar to the moving average method with the difference that allows different weights to be assigned to the previous values. The idea is that the most recent values should have more weight in the forecast value. The forecast made for time $t$, $\hat{y}_t$, by the weighted moving average model is obtained as follows:

$$\hat{y}_t = \sum_{i=1}^{n} w_{t-i} \cdot y_{t-i} \tag{2.8}$$

where $w_{t-i}$ is the weight given to the value of time $t-i$ and such that $\sum_{i=1}^{n} w_{t-i} = 1$.

#### 2.3.1.4 Exponential Smoothing

The exponential smoothing can be seen as a refined weighted moving average method, since it also gives weight to the past values but using a smoothing constant for that purpose.

This approach can be divided in three different methods and the selection of the method is generally based on the components of the time series (trend and seasonality).

**Simple Exponential Smoothing -** This first method for exponential smoothing is the most simple one, hence the name Simple Exponential Smoothing (SES). This method is suitable for time series that show no obvious trend or seasonality. Assuming that the series starts at time 0, the simplest form of exponential smoothing is given by the recurrence:

$$\begin{aligned} \hat{y}_0 &= y_0 \\ \hat{y}_t &= \alpha \cdot y_t + (1 - \alpha) \cdot \hat{y}_{t-1} \end{aligned} \tag{2.9}$$

where $0 \leq \alpha \leq 1$ is the smoothing constant. Values of $\alpha$ closer to zero will give a greater smoothing effect and turn the method less responsive to recent changes. Values of $\alpha$ closer to one will give more weight to recent changes in the data and a lower smoothing effect. Thus, the forecast for time $t$ is a weighted average of all the observations in the time series. The weights given to older observations decrease faster or slower according to the parameter $\alpha$.

**Holt's method -** In order to forecast time series with trend, Holt (2004) extended simple exponential smoothing and created Holt's method, also called double exponential smoothing since it has two smoothing equations, one for the level and another for the trend.

**Holt-Winters method -** Similary to the Holt's method that extends SES method for series with trend, Holt (2004) and Winters (1960) have extended the Holt's method to create Holt-Winters and capture seasonality. Just like Holt's, the Holt-Winters adds another smoothing equation to the previous method, in this case the seasonality component, having now three smoothing equations, hence the name "Triple Exponential Smoothing".

#### 2.3.1.5 Autoregressive Integrated Moving Average (ARIMA)

The Auto-Regression Integrated Moving Average (ARIMA) model was developed in the 1970's by George Box and Gwilym Jenkins, also being called “Box-Jenkins model”. This model predicts the future values in a time series based on three parameters:

$p$ **-** number of autoregressive terms (AR);

$d$ **-** number of nonseasonal differences (I);

$q$ **-** number of moving-average terms (MA).

To apply this method we need to differencing the series a couple of $d$ times until we have a completely stationary series. After that, we still need to determine the $p$ and $q$ values, and for that, we can try different combinations and see what works best. We can also define them by using ACF plot using techniques that allow us to identify the best $p$ and $q$ values.

In order to automate ARIMA, a function `auto.arima` has been developed by Hyndman et al. (2007). This function takes into account the Akaike Information Criterion (AIC) to compare several ARIMA models. The parameters of ARIMA are then chosen by minimizing the AIC. For this purpose, the algorithm executes a stepwise search to select the ARIMA model with the smallest AIC.

#### 2.3.1.6 TBATS

The Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality (TBATS) model proposed by De Livera et al. (2011) tried to do a new approach to solve specific problems in the forecast task like complex seasonal patterns, such as multiple seasonal periods, or high-frequency seasonality. TBATS model can be defined and written as:

$$TBATS(\omega, \{p, q\}, \varphi, \{< m_1, k_1 >, < m_2, k_2 >, \cdots, < m_T, k_T >\}) \tag{2.10}$$

where:

$\omega$ is a Box-Cox transformation (Box and Cox, 1964),

$p$,$q$ are ARIMA parameters (Box and Jenkins, 1976),

$\varphi$ is a damping parameter (Gardner Jr and McKenzie, 1985),

$m_1, \cdots, m_T$ are seasonal periods,

$k_1, \cdots, k_T$ are are number of Fourier series pairs (West and Harrison, 2006).

This model differs from ARIMA and Holt-Winters since it can deal with dual calendar and high frequency multiple seasonal patterns much better, allowing the seasonality to change slowly over the time in the prediction model.

### 2.3.2 Non-Linear Models

Nowadays, complex time series require the use of more advanced methods, such as non-linear models, that unlike linear models, allows the use of several explanatory variables. These variables will increase the accuracy and give better understanding of the behavior of the target variable. Nevertheless, it is always necessary to be careful when adding a explanatory variable to not add unintentional noise to the data.

#### 2.3.2.1 Artificial Neural Networks

The Artificial Neural Network (ANN) is a machine learning model that mimics the behaviour of how natural biological neural networks work. Just as a neuron receives several inputs and produces a single output to another neuron, we will do this process to create a neural network. The creation of a neural network starts with building a perceptron that can be seen as the most basic form of a neural network, a simple binary function that only has two possible results. The perceptron receives inputs, multiplies them by some weights (usually the weights are randomly generated) and sends as input to an activation function that will produce the binary output. The perceptron also has a bias to make sure if all inputs are equal to zero there is not going to be any issues, since the multiplicative weight would not have any effect. The activation function can be trigonometric function, step function, etc. After the first output, we start to adjust the weights and repeat the process until we reach the maximum number of iterations or an acceptable error rate.

A ANN model consists of the combination of layers of perceptrons together. The input layer receives the initial inputs. The output layer outputs the results of the ANN. In the middle of these two layers, we have the hidden layers, they are called this way because the perceptrons in these layers never directly see the initial inputs or the final output (Portilla). An example of an ANN can be seen on the Figure 2.6.

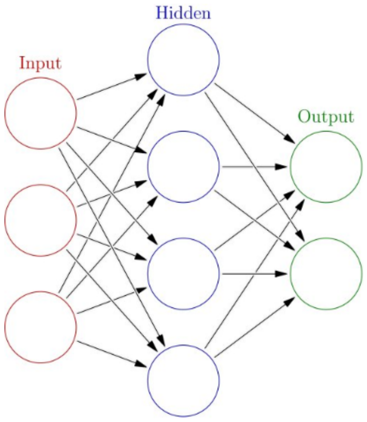


Figure 2.6: Diagram of a Neural Network (Portilla)

Neural networls have been widely used as a method for time series forecasting, for instance in many business and economic time series with seasonal and trend variatons. One of the most important advantages is that no assumptions need to be taken about the data, since all the relationships are determined through data mining, to model complex nonlinear relationships (Zhang and Qi, 2005). The ANN also showed to be able to capture the nonlinear trend and seasonal patterns and the interactions between them from the time series (Alon et al., 2001).

#### 2.3.2.2 Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression problems. Given a training data, this algorithm outputs an optimal line or hyper-plane in multidimensional space, that gives the largest minimum distance between the training examples. This line should not pass to close to the sample points to avoid noise in the data and possible interference in the final prediction, so the goal is always to find a line that is as far away as possible from the observations (OpenCV). An example of a hyper-plane can be seen on the Figure 2.7.

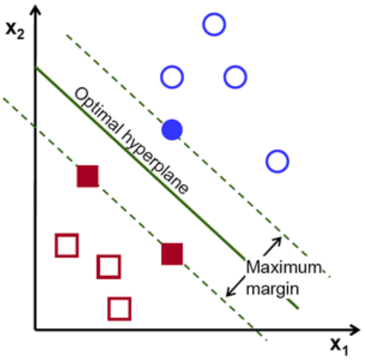


Figure 2.7: An optimal Hyper-Plane (OpenCV)

Still, most of the real life problems are not linearly separable so we have to project the data into a higher dimensional space, for this we apply the kernel trick. An application of the kernel function can be seen in the Figure 2.8

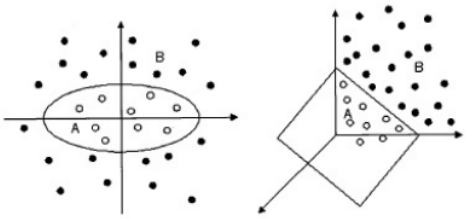


Figure 2.8: Mapping of non-linear separable training data from $\mathbb{R}^2$ into $\mathbb{R}^3$ (Hofmann, 2006)

In He et al. (2014), the SVM was used to forecast a time series of the river flower while comparing the results with other methods, such as ANN. In the end, it was conclude that the SVM did a better job, although all achieved good results.

One of the limitations behind SVM is that the results are very sensitive to the parameters, like for example, upper bound C and the kernel parameter, making it crucial to find the optimal parameters (Kim, 2003).

### 2.3.3 Ensemble Models

Ensemble model is a machine learning technique that consists on having a set of prediction models that combine each of the individual decisions to create one optimal predictive model. Several studies (e.g. Dietterich (2000)) have shown that, when compared to the individual prediction models present in the ensemble, ensemble models have a higher accuracy rate. Even though ensemble models exponentially increase the executing time and consequently need more computational power, the benefits we can gain from it are much higher, since it reduces the error and avoids overfitting.

#### 2.3.3.1 Random Forests

A commonly used ensemble technique is the Random Forest algorithm (Breiman, 2001) that can be considered a parallel ensemble method since the base learners are executed in parallel. The parallel execution allows to run the different base learners independently increasing the accuracy by using the average results (Hall et al., 2003). Random Forest is a predictor that generates an ensemble of M decision trees during training phase (Biau and Scornet, 2016). These trees are built using a bootstrap sample of the target data and a branch of the tree represents a possible decision. At each node the data is split and optimal variables are selected from a random subset of the available variables, this guarantees a diversity and randomness while growing the tree, resulting in a better model. The final prediction of the random forest is the majority decision of all the trees (Degenhardt et al., 2017).

Kane et al. (2014) performed a comparation between the Random Forests model and the ARIMA model when forecasting data of outbreaks of highly pathogenic avian influenza (H5N1) in Egypt. The Random Forest model was able to outperform ARIMA due to the ability of incorporating non-linear relationships into the forecast task.

#### 2.3.3.2 Tsensembler

The goal of Tsensembler (Cerqueira et al., 2017) is to combine several forecasting models applied to a time series using metalearning and other methods of combining predictions.

The combination strategies usually involve analyzing the error of the previously chosen methods and adapting the weight to each of them. Another common approach is stacking (metalearning approach). As mentioned by Wolpert (1992), this strategy does not follow the winner-takes-all philosophy, meaning that it does not simply choose the learning method with the smallest error of prediction, but instead it incorporates the outputs of different learning models in a new data set that is then treated as a new prediction problem.

The metalearning strategy used by Tsensembler, named Arbitrated Dynamic Ensemble (ADE), unlike the stacking method, focuses on the importance of giving more weight to forecasting methods that are more appropriated to the time series. For this to be possible, for each base-learner (prediction method) a meta-learner is built, which allows to estimate how suitable is the respective base-learner to predict the target time series. While all base-learners $M$ produce their predictions, meta-learners $Z$ compute the $W$ weights taking into account the calculated prediction errors ($\widehat{e}_i$). The final output of ADE is given using a weighted average of the predictions relative to the weights. All this work flow can be observed in the Figure 2.9.

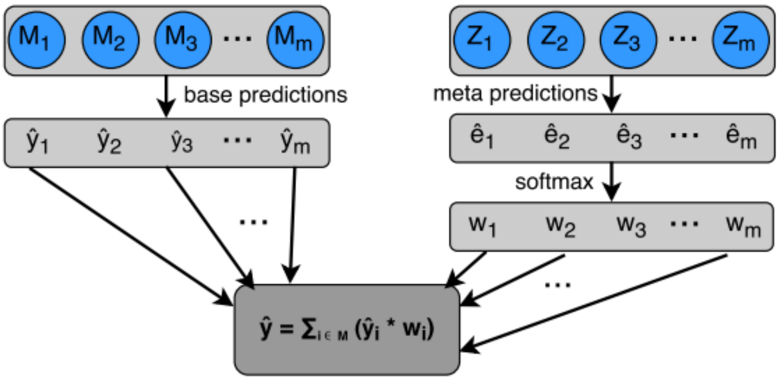


Figure 2.9: Scheme of the workflow of ADE for a new prediction (Cerqueira et al., 2017)

### 2.3.4 Hierarchical Models

The hierarchical models are very popular for forecasting time series in domains where data has a natural hierarchical structure. Examples of such structure can be found in several domains, such as biology (e.g. species follow a hierarchical taxonomy), chemistry (e.g. pharmaceuticals can be grouped by their purpose, like relieve pain or help with feelings of depression) or retail industry (e.g. appliances such as refrigerators or washing machines can be grouped together).

In Figure 2.10 we can see a basic hierarchical structure with two levels. In this structure, the level 0 (Total) represents the aggregation of all the existent series. The level 1 divides the data in three different groups (A, B and C). desegregating the series, and so on and so forth. The level 2 disaggregates each of the previous groups in three groups more, obtaining nine different series. This disaggregation process can continue until the maximum disaggregation of the series is reached.

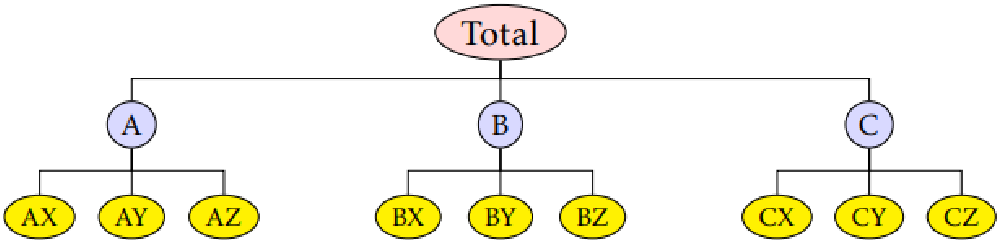


Figure 2.10: Hierarchical Time Series (Hyndman)

At any time $t$, any observation from the bottom level of the hierarchy is summed up into the upper levels. The hierarchy presented in Figure 2.10 can be represented by the following summing matrix:

$$\begin{pmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \\ y_{AX,t} \\ y_{AY,t} \\ y_{AZ,t} \\ y_{BX,t} \\ y_{BY,t} \\ y_{BZ,t} \\ y_{CX,t} \\ y_{CY,t} \\ y_{CZ,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_{AX,t} \\ y_{AY,t} \\ y_{AZ,t} \\ y_{BX,t} \\ y_{BY,t} \\ y_{BZ,t} \\ y_{CX,t} \\ y_{CY,t} \\ y_{CZ,t} \end{pmatrix} \tag{2.11}$$

The observed value of the series at time $t$, $y_t$, is given by:

$$y_t = S \cdot y_{K,t} \tag{2.12}$$

where $S$ is the summing matrix presented above and $y_{K,t}$ is a vector with the bottom level observations of the hierarchy at time $t$.

There are three main prediction methods commonly used in this hierarchical series: bottom-up, top-down and the combination of the two.

#### 2.3.4.1 Bottom-up method

Bottom-up is the most used method for hierarchical prediction. This technique starts by independently forecasting every bottom level series and then proceeds by aggregating the results of the forecast to produce forecasts to the upper level. For example, for the hierarchy presented in Figure 2.10, to produce the $h$-step-head forecast of the group A in the level 1, we need to sum up the base forecast of the bottom level series:

$$\hat{y}_{A,h} = \hat{y}_{AX,h} + \hat{y}_{AY,h} + \hat{y}_{AZ,h} \tag{2.13}$$

where the $\hat{y}_{A,h}$ is the $h$-step-head forecast for the series A.

These forecasts are aggregated using the summing matrix S, this way the bottom-up approach can be defined and written as:

$$\hat{y}_h = S \cdot \hat{y}_{K,h} \tag{2.14}$$

where $\hat{y}_h$ is the vector with all the final forecasts of the hierarchy.

This approach has the advantage of not loosing any data, since the base forecast are made at the most disaggregated level of the hierarchy. However, it should be noticed that the bottom level is the level with more noise, which is going to influence the forecast of the hierarchy.

#### 2.3.4.2 Top-down method

The top-down approach begins by making the forecast for the top level series (Total in Figure 2.10) and then disaggregates it downward to the rest of the hierarchy. For this purpose uses a set of proportions, $p_1$ ... $p_m$ , that define how the base forecast is going to disaggregated for each series at the bottom level of hierarchy.

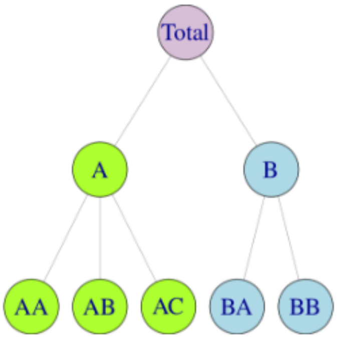


Figure 2.11: Example of a two level hierarchical tree (Rob J Hyndman).

For example, for the hierarchy in the Figure 2.11 using the proportions $p_1$, ... , $p_5$ we get:

$$\tilde{y}_{AA,t} = p_1\hat{y}_t \tag{2.15}$$

$$\tilde{y}_{AB,t} = p_2\hat{y}_t \tag{2.16}$$

$$\tilde{y}_{AC,t} = p_3\hat{y}_t \tag{2.17}$$

$$\tilde{y}_{BA,t} = p_4\hat{y}_t \tag{2.18}$$

$$\tilde{y}_{BB,t} = p_5\hat{y}_t \tag{2.19}$$

After the bottom-level h-step-ahead forecasts are calculated, they method starts to aggregate the results to generate forecasts for the rest of the hierarchy. The top-down approach can be represented by the following equation:

$$\tilde{y}_h = S_{p\hat{y}_t} \tag{2.20}$$

#### 2.3.4.3 Middle-out approach

In the middle-out approach the bottom-up and the top-down approaches are combined. It starts by choosing a middle level and generating base forecasts for all the series in it. After that, to calculate the forecasts for upper levels, the bottom-up approach is used, and to determine the lower level forecasts, the top-down approach is used. The bottom-up method aggregates the middle base forecasts, while the top-down disaggregates the middle base forecasts to generate the forecasts for the rest of the hierarchy.

#### 2.3.4.4 Optimal forecast combination

Hyndman et al. (2015) has shown that, assuming the prediction errors have the same aggregation constrains as the data, it is possible to obtain predictions for each level of the hierarchy by using the equation:

$$\widetilde{Y}_t(h) = S(S^t S)^{-1} S^t * \widehat{Y}_t(h) \tag{2.21}$$

where $\widetilde{Y}_t(h)$ represents the revised forecast for the series $h$ in time $t$, and $\widehat{Y}_t(h)$ represents the previous obtained forecast for the series $h$ in time $t$.

### 2.3.5 Spatial-Temporal Models

Spatial-Temporal models have been used for several decades, filling a need to map behaviours and patterns, and recognized hidden correlations between the data (Christakos, 2000). This models are proposed when there is both spatial and temporal components in the data set, and are very popular in many areas like hydrology and ecology.

Conley et al. (2008) used the Spearman correlation and the two-way ANOVA (Analysis of Variance) to test the relationships between the concentration of a pharmaceutical and different spatial (depth, Sewage Treatment Plant proximity, and downstream distance from headwaters) and temporal (seasonality) factors. Other examples of these models are analysis on rainfall data and the risk of flood (Wheater et al., 2005), and on air pollution affecting human health (Warren et al., 2012).

### 2.3.6 Evaluation of Forecasting Models

When building a forecasting model, we can only estimate its performance if we test it against unseen data. A test over the data used to train the model would give us an optimistic estimate of its performance. Thus, to obtain a more realistic estimate of the model's performance, we split the data set into two parts, creating the training and test data. This way, we can use the training data to train the model and the rest of the data, the test data, to test our accuracy. A common practice among the statisticians is the training data having around 80% of the sample while the test data get about 20%. However, this is not always possible and relies heavily on the sample size. Moreover, we should be aware that increasing too much the training data size can lead to over-fitting making it unable to make good predictions except for that same data.

A possible approach to split the data set is separating the data in two different windows, before and after observations at time $t$. For this job we can have three different alternatives for the learning task:

**fixed Window -** when there is a fixed number of consecutive observations used for training (possibly all the available training data) and the obtained model is used to make all the predictions for the observations in the test set;

**sliding Window -** if for each observation or group of observations in the test set, a new model is created based on a fixed number of past observations;

**growing Window -** similar to the sliding window technique but, instead of only using a fixed number of past observations, it uses all the data available till then.

#### 2.3.6.1 Monte Carlo Simulation

One of the main characteristics of time series data sets is that each observation has a "time" attribute associated with it, which establish an order between them. During the various data mining steps, such as pre-processing, choosing model and later predicting future data, it is always necessary to take into account and respect the order of the data with the risk of not taking account the time factor and obtain inaccurate results. With this, we know that any method involving random re-sampling that changes the order of the data cannot be used to obtain reliable estimates for the evaluation metrics, since we could not guarantee that we would not test on observations that are older than the ones used for training the model.

Monte Carlo experiment successfully allows to randomly test our data and ensure statistical reliability while respecting the time ordering. Given a time series $(t)$ of $n$ length, we first need to choose the size of both train $(w_{train})$ and test window $(w_{test})$, always keeping in mind that the sum of both windows needs to be much smaller than $n$ so that we can create different scenarios of training and testing in the time series. Now that the dimensions of both windows are defined, we can randomly generate multiple time points $(r)$, and for each one of them we will use the data in the interval $[r - w_{train}, r]$ to train and then the interval $]r, r + w_{test}]$ to test. It is important to note that while randomly choosing the time points $r$, we need to make sure that there is always enough data from both past and future observations to guarantee the length of the train and test window (Torgo, 2016).

#### 2.3.6.2 Error metrics

The best way to evaluate the performance of the various models of prediction is, without doubt, using an error measure. In theory, the model that shows in most of the cases, a lower error value it should be the most appropriated for that problem. So, for our prediction problem we selected Mean Absolute Error (MAE) as our measure error. MAE is the mean of the absolute values of the differences between the forecast value ($\hat{y}_i$) and the actual values ($y_i$).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2.22}$$

MAE gives the accuracy in the same units as the data so the smaller the value, the more accurate the forecast. However, we cannot compare MAE values from different data set because they depended always on the magnitude of the data.

For this reason, Mean Absolute Percentage Error (MAPE) is one of the alternatives commonly used to measure the forecast error. This metric normalizes the MAE as follows,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100 \tag{2.23}$$

Still, it should be noticed that this metric cannot be used in time series with zeros, as the percentage cannot be calculated for such cases.

## 2.4 Tools

R (Team et al. (2013)) is a programming language and an environment that can be used for statistical computing, generate graphics, analysis and manipulation of data. Created as a GNU project and identical to the S language created at Bell Laboratories, is many times considered as a different implementation of S. R was developed in the year of 1996 by Ihaka and Gentlement in the University of Auckland and nowadays, taking advantage of its open source policy, the current development is in the hands of a not so big group, from different countries, and the growing community itself.

R gives a wide range of different tools that allows us to do classification, clustering, linear modeling, time-series analysis and many more statistical methods. To demonstrate the results of the methods, R can easily generate graphics or tables that can be exported to use, for example, in a paper or presentation.

R has several packages that provide functions that helped us in many ways. Some of the packages we used for this study were the following. The package `xlsx` (Dragulescu et al., 2018) to import excel files to our R enviroment. The package `forecast` (Hyndman et al., 2007) to create the time series objects and provided several prediction models that we used in this work.

For the ARIMA model we used the package stats. The packages e1071 (Meyer et al., 2018), rpart (Therneau et al., 2010) and randomForest (Liaw et al., 2002) to generate the SVM, rpart and randomForest models, respectively. tsensembler (Cerqueira et al., 2017) to create the embedded series and apply the Arbitrated Dynamic Ensemble method. For the creation and forecasting of the hierarchical models we used the package hts (Hyndman et al., 2015). The package parallel (Team, 2013) allowed us to use multiple cores, which reduced significantly the time we needed for retrieving information and run some of the prediction models. The package dataRetrieval (Hirsch and De Cicco, 2015) was used to get the data from the USGS Portal. In order to generate all the plots for this thesis, we used the package ggplot2 (Wickham, 2010). We chose Rstudio (Racine, 2012) to be our work environment for the R language.

# Chapter 3

# Case Study 1: Forecasting Pharmaceuticals Pollutants in Water

Our first case study concerns the prediction of pharmaceuticals pollutants in a river in Portugal. The lack of knowledge and the uncertainty about the impact these emerging pollutants can have in the human health make this a relevant topic, specially if we add to the problem that the Wastewater Treatment Plants (WWTPs) are not ready to deal with this type of contaminants. It is even worse if we think that, every day, even more compounds are released by industries, agriculture and waste waters from population. The goal of this case study is to try to understand the behaviour of this new chemicals, which are still not regulated by any directive, in order to be able to forecast their concentration values.

## 3.1 Data set

The data consists of a set of, approximately, monthly observations from August 2013 to June 2014 collected in several locations of Rio Lis measuring the concentration levels of previously chosen pharmaceuticals. The samples were collected at Rio Lis, located in the central region of Portugal, due to its known high level of pollution. Pollution can be partly explained by the fact that it is a calcareous area, meaning that it is a zone that favors water infiltration. As there are not enough WWTPs to cover the whole population around the river, some untreated water reaches the river ecosystems leading to low quality of drinking water. It is also important to reinforce that water contamination, does not come exclusively from population but also from hospital effluents, piggeries effluents and landfill leachate. Even though some of these external sources go through Wastewater Treatment Plant (WWTP), before reaching the river, not all the pharmaceuticals waste gets filter, ending up harming the ecosystem and the quality of the drinking water (Paíga et al., 2016).

With a length of $39.5km$ and covering about $495km^2$, the river has two WWTPs, Olhavas and Coimbrão, at a distance of 25km and 5km from river mouth, respectively.

Nine sampling points were selected: five along the river and four in the two WWTPs. On the river samples were taken in the river spring, and upstream and downstream of each WWTP. While in the WWTPs, it was on the influent and effluent for each one.

In Figure 3.1 we can observe all nine different locations where the samples were obtained. Still, it is important to note that, in the figure, the influent and effluent of both WWTPs are represented together. WWTP1 represents Olhavas WWTP and WWTP2 represents Coimbrão WWTP.

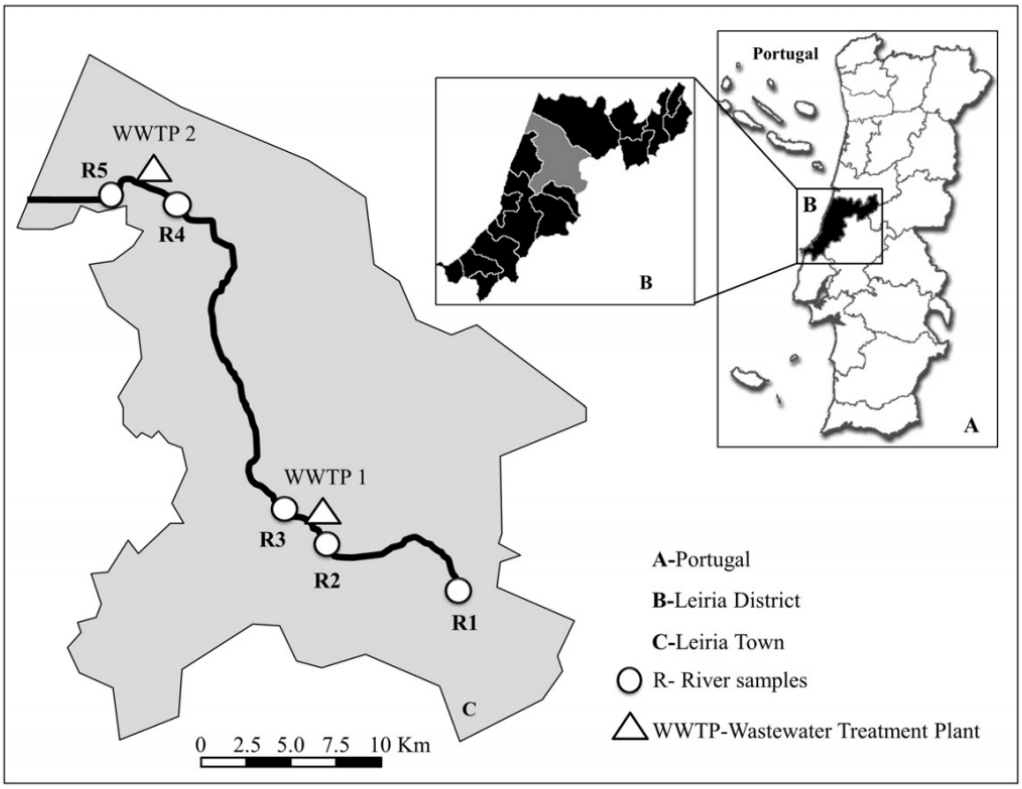


Figure 3.1: Sampling Locations of Case Study 1 (Paíga et al., 2016).

To simplify the designation of these sampling locations, we adopted the names presented in Table 3.1.

For each one of these sampling locations it was registered the concentration of 32 pharmaceuticals that can be organized in a 4 level hierarchy as shown in Figure 3.2.

The first level aggregates all the pharmaceuticals considered in this work. Just below, the pharmaceuticals are divided into three groups:

- **antibiotics**, which aim to kill bacteria and prevent their future appearance; they are used to treat some types of infections and they need to be prescribed by a doctor.

| Designation | Sampling Location |
|---|---|
| NASCEN | River Spring |
| ETAROM | Upstream of Olhavas WWTP |
| ETAROE | Influent of Olhavas WWTP |
| ETAROS | Effluent of Olhavas WWTP |
| ETAROJ | Downstream of Olhavas WWTP |
| ETARNM | Upstream of Coimbrão WWTP |
| ETARNE | Influent of Coimbrão WWTP |
| ETARNS | Effluent of Coimbrão WWTP |
| ETARNJ | Downstream of Coimbrão WWTP |

Table 3.1: Designations of the Sampling Locations for Case Study 1.

- **psychiatric drugs**, which try to lessen or remove symptoms caused by depression, usually in conjunction with therapy; this is achieved by increasing the chemicals called neurotransmitters, which manage to improve a person's mood and in some cases help overcome long-term pain;
- **non-steroidal anti-inflammatory drugs (NSAIDs) / analgesics**, which help to block certain chemicals that cause inflammation and are also efficient to treat general or localized pain.

Before the last level of the 32 pharmaceuticals, we still have 11 more subgroups that, for example, divide the pharmaceuticals belonging to the group of antibiotics in Macrolides and Sulfonamides. This hierarchy allows a better understanding of the similarity in the behavior of some drugs. The complete hierarchical structure of the considered pharmaceuticals is shown in Figure 3.2.

For each location in each day of sampling, two tests were carried out and each of them with two observations. Then the mean of these two tests was calculated, resulting in the value of the concentration obtained through the chemical analysis technique LC-MS/MS (Liquid Chromatography - Mass Spectrometry). As the last step, the correction of the recovery was applied to the previous value, being this value the final one that was used for analysis and forecast in this work.

Attached to the observations of the pharmaceuticals concentration values, there is a table with the values of Method Detection Limit (MDL) and Method Quantitation Limit (MQL) of each pharmaceutical for the observations in the river and in the influent and effluent of the WWTP. All the cases in which the presence of a pharmaceutical was detected but does not reach the limit imposed by the analysis method, it was not possible to assign a value of concentration. For those observations the concentration value is recorded as "<MDL" or as "<MQL" in the database. In the case that the pharmaceutical was not detected at all, his concentration value is recorded as "n.d.".

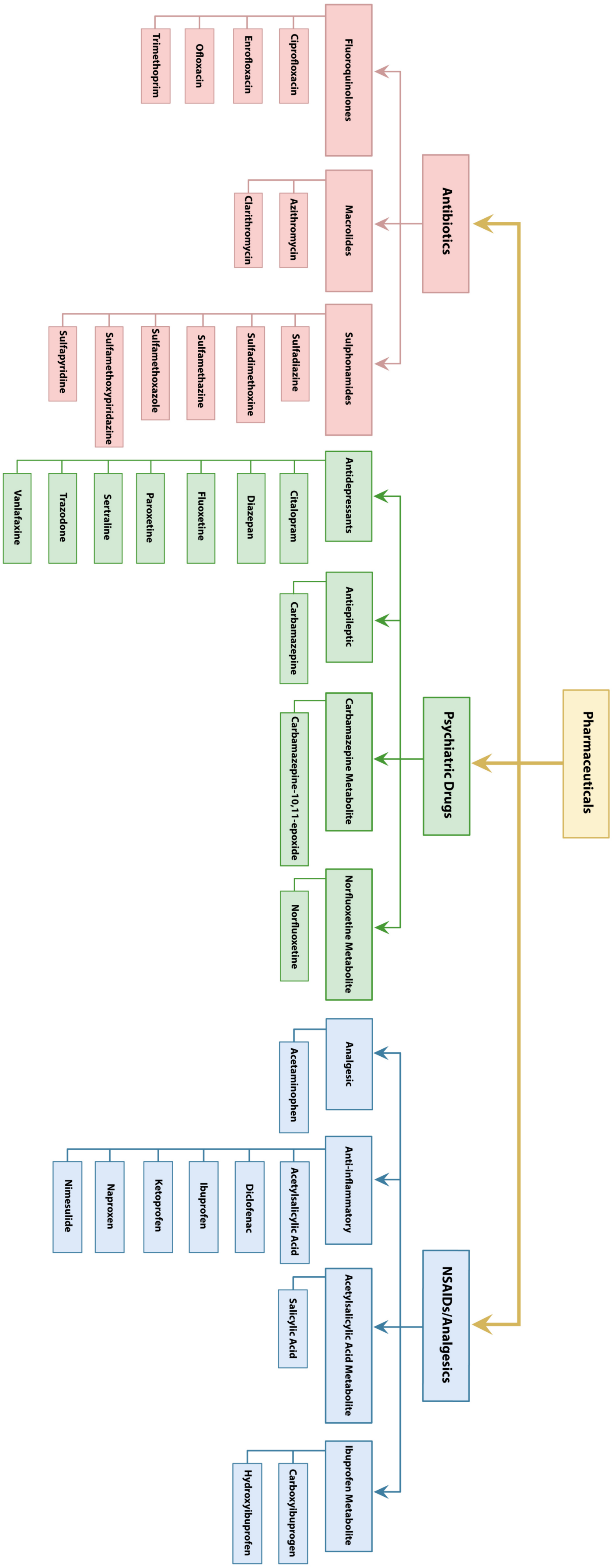


Figure 3.2: Pharmaceutical Hierarchy

## 3.2 Data Analysis and Pre-Processing

The original data set was initially composed by the following four attributes:

**location:** one of the nine different sampling points across the river or in one of the two WWTP;

**date:** day, month and year of the sampling;

**pharmaceutical:** the pharmaceutical for which the concentration was tested.

**concentration:** value that can be numerical, or categorical if the value did not reach the minimum limit of the detection method in which case is registered as "<MQL".

To this data set we added the following three new attributes:

**season:** the season of the year; it might be useful to aid us in the analysis of the seasonality and, possibly, in the prediction of values in specific seasons;

**group:** the group of the pharmaceutical;

**subgroup:** the subgroup of the pharmaceutical.

The samples comprise a time window from August 2013 to June 2014, with a total of 2912 observations of the concentrations corresponding to 91 samples taken from the river and WWTP. Six drugs were never detected, namely Acetylsalicylic Acid, Enrofloxacin, Nimesulide, Ofloxacin, Sulfadimethoxine and Sulfamethoxypiridazine. In addition to this six drugs that were never detected, there are also two drugs, Sertraline and Sulfadiazine, which although they were detected, they were always below the limits of the methods used for detection. Thus, the observations related to these two drugs were removed from the data set. All the remaining values not detected or below the detection limit were replaced by zero or by the value of the limit, respectively.

The number of observations of the Antibiotics group is slightly larger than the other two, Psychiatric drugs and the NSAIDs/Analgesics, both with the same number of observations (cf. Figure 3.3). The reason for this is that the Antibiotics group has 12 drugs and not 10 as the other two. Despite having the highest number of observations, the Antibiotics group stands out for having a much lower number of detected drugs, reaching a value of 8.15% of detected rate, while the other two groups, Psychiatric drugs and the NSAIDs/Analgesics, reach higher percentages, 37.47% and 51.87%, respectively. The difference of detected rates between the pharmaceuticals groups can be seen in the heat maps (figs. 3.4 to 3.6), one for each group, that show the detected percentage for each pharmaceutical for every sampling location.

When we start to analyze the value of the concentrations that were detected, we can observe that the scenario is different: the group of Antibiotics no longer stands out, and now we have the group of NSAIDs/Analgesics showing very high levels of concentration when compared to the rest.

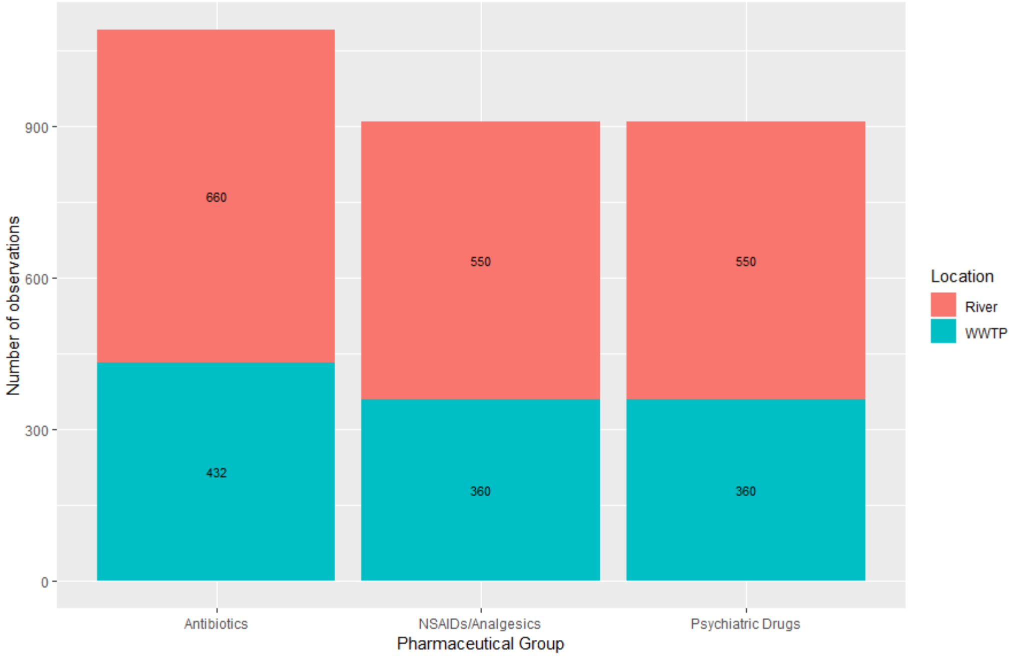


Figure 3.3: Number of observations for each pharmaceutical group.

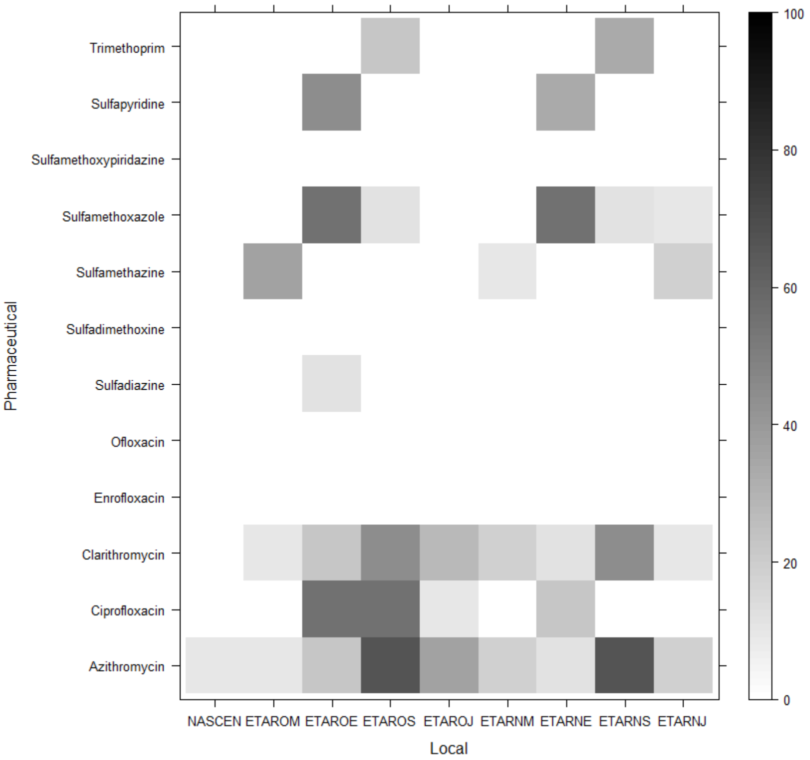


Figure 3.4: Heat map of the detection rate for pharmaceuticals in the Antibiotics group.

Figure 3.7 presents the concentration values of the three groups along the various sampling sites. Due to the problem of the high values of the group NSAIDs/Analgesics, it was necessary to present also the logarithmic scale transformation, otherwise it would not be possible to distinguish

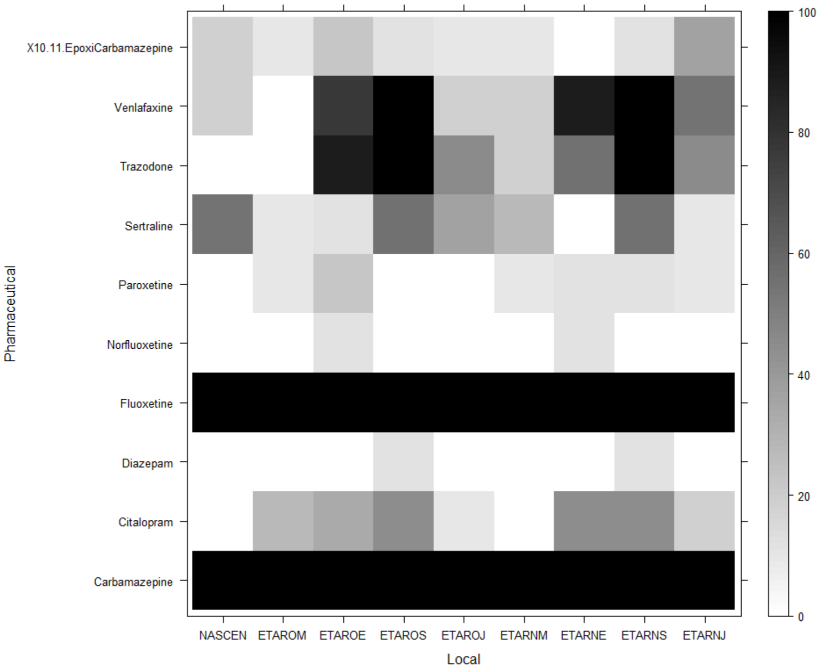


Figure 3.5: Heat map of the detection rate for pharmaceuticals in the Psychiatric Drugs group.

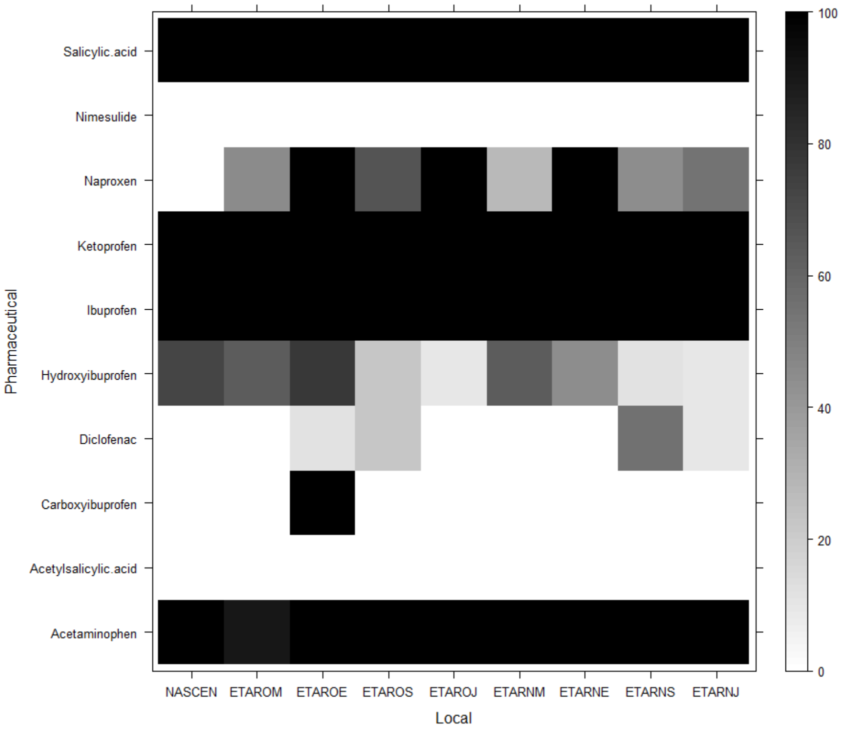
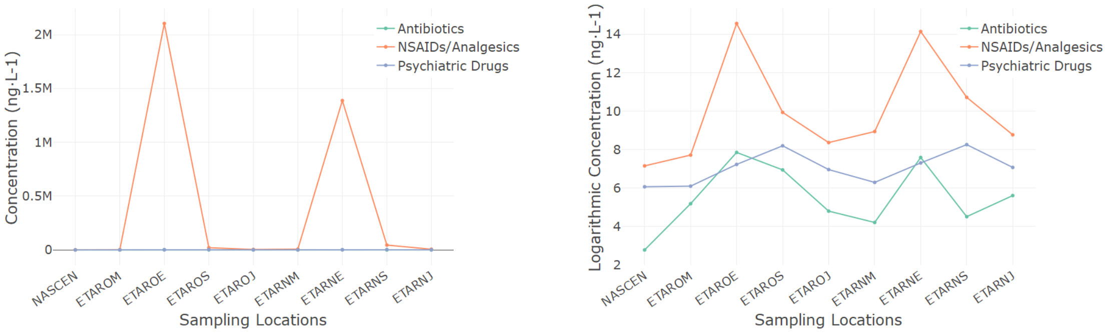


Figure 3.6: Heat map of the detection rate for pharmaceuticals in the Analgesics group.

the two other groups.

As we can see by Figure 3.7b and by Table 3.2, the NSAIDs/Analgesics group despite having



(a) original scale

(b) logarithmic scale

Figure 3.7: Concentration for each group of drug observed in all the sampling locations.

the highest values found in the influent of both WWTP, shows the highest removal efficiency. This removal efficiency for each pharmaceutical was evaluated using the Equation 3.1.

$$\text{Removal efficiency (\%)} = \frac{(c_{inf} - c_{eff})}{c_{inf}} \tag{3.1}$$

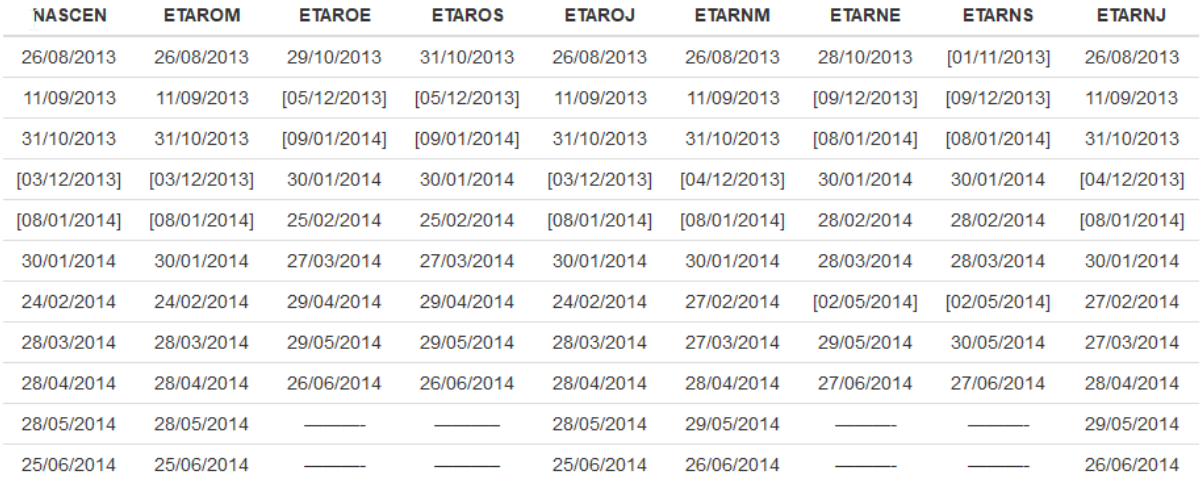
where $c_{inf}$ and $c_{eff}$ represent the concentration found in the influent and effluent of the WWTP, respectively.

Two of the main reasons for these removal efficiency values are the high degradation rate of this kind of drugs and the fact that their removal is independent of the hydraulic retention time (amount of time that the water remains storage in the WWTP).

| | Olhavas's WWTP<br>(WWTP1) | Coimbrão's WWTP<br>(WWTP2) |
|---|---|---|
| Antibiotics | 59.69% | 95.46% |
| Psychiatric Drugs | -163.37% | -159.43% |
| NSAIDs/Analgesics | 99.02% | 96.74% |

Table 3.2: Removal efficiency of the drug groups in the two WWTPs.

Irregularity in time series is frequently in systems not automated for collecting samples and where human intervention is needed. This is the case of our data set, where we have discrepancies in the time lag between the sampling dates. In this type of time series, as the observations are not equally spaced in time, finding patterns becomes harder to some prediction models. Therefore, what we decided to do first was to discard the day of the sample and only use the month and the year, assuming that the period between each sampling was monthly. Afterwards, we realized that only using the month brought another problem, because not every month has an observation. For example, in the river spring there is no sample in November, but we have two samples in January. One workaround to this problem is changing the month of a few sampling dates, so that we have a sample every month from August of 2013 to June of 2014 in the river spring. This option was taken after contacting with the researchers of the project who told us that the initial idea was doing monthly samples, so it was acceptable to make this modification to our data set. Each date that has been modified is indicated with brackets in Figure 3.8.



| NASCEN | ETAROM | ETAROE | ETAROS | ETAROJ | ETARNM | ETARNE | ETARNS | ETARNJ |
|---|---|---|---|---|---|---|---|---|
| 26/08/2013 | 26/08/2013 | 29/10/2013 | 31/10/2013 | 26/08/2013 | 26/08/2013 | 28/10/2013 | [01/11/2013] | 26/08/2013 |
| 11/09/2013 | 11/09/2013 | [05/12/2013] | [05/12/2013] | 11/09/2013 | 11/09/2013 | [09/12/2013] | [09/12/2013] | 11/09/2013 |
| 31/10/2013 | 31/10/2013 | [09/01/2014] | [09/01/2014] | 31/10/2013 | 31/10/2013 | [08/01/2014] | [08/01/2014] | 31/10/2013 |
| [03/12/2013] | [03/12/2013] | 30/01/2014 | 30/01/2014 | [03/12/2013] | [04/12/2013] | 30/01/2014 | 30/01/2014 | [04/12/2013] |
| [08/01/2014] | [08/01/2014] | 25/02/2014 | 25/02/2014 | [08/01/2014] | [08/01/2014] | 28/02/2014 | 28/02/2014 | [08/01/2014] |
| 30/01/2014 | 30/01/2014 | 27/03/2014 | 27/03/2014 | 30/01/2014 | 30/01/2014 | 28/03/2014 | 28/03/2014 | 30/01/2014 |
| 24/02/2014 | 24/02/2014 | 29/04/2014 | 29/04/2014 | 24/02/2014 | 27/02/2014 | [02/05/2014] | [02/05/2014] | 27/02/2014 |
| 28/03/2014 | 28/03/2014 | 29/05/2014 | 29/05/2014 | 28/03/2014 | 27/03/2014 | 29/05/2014 | 30/05/2014 | 27/03/2014 |
| 28/04/2014 | 28/04/2014 | 26/06/2014 | 26/06/2014 | 28/04/2014 | 28/04/2014 | 27/06/2014 | 27/06/2014 | 28/04/2014 |
| 28/05/2014 | 28/05/2014 | ——— | ——— | 28/05/2014 | 29/05/2014 | ——— | ——— | 29/05/2014 |
| 25/06/2014 | 25/06/2014 | ——— | ——— | 25/06/2014 | 26/06/2014 | ——— | ——— | 26/06/2014 |

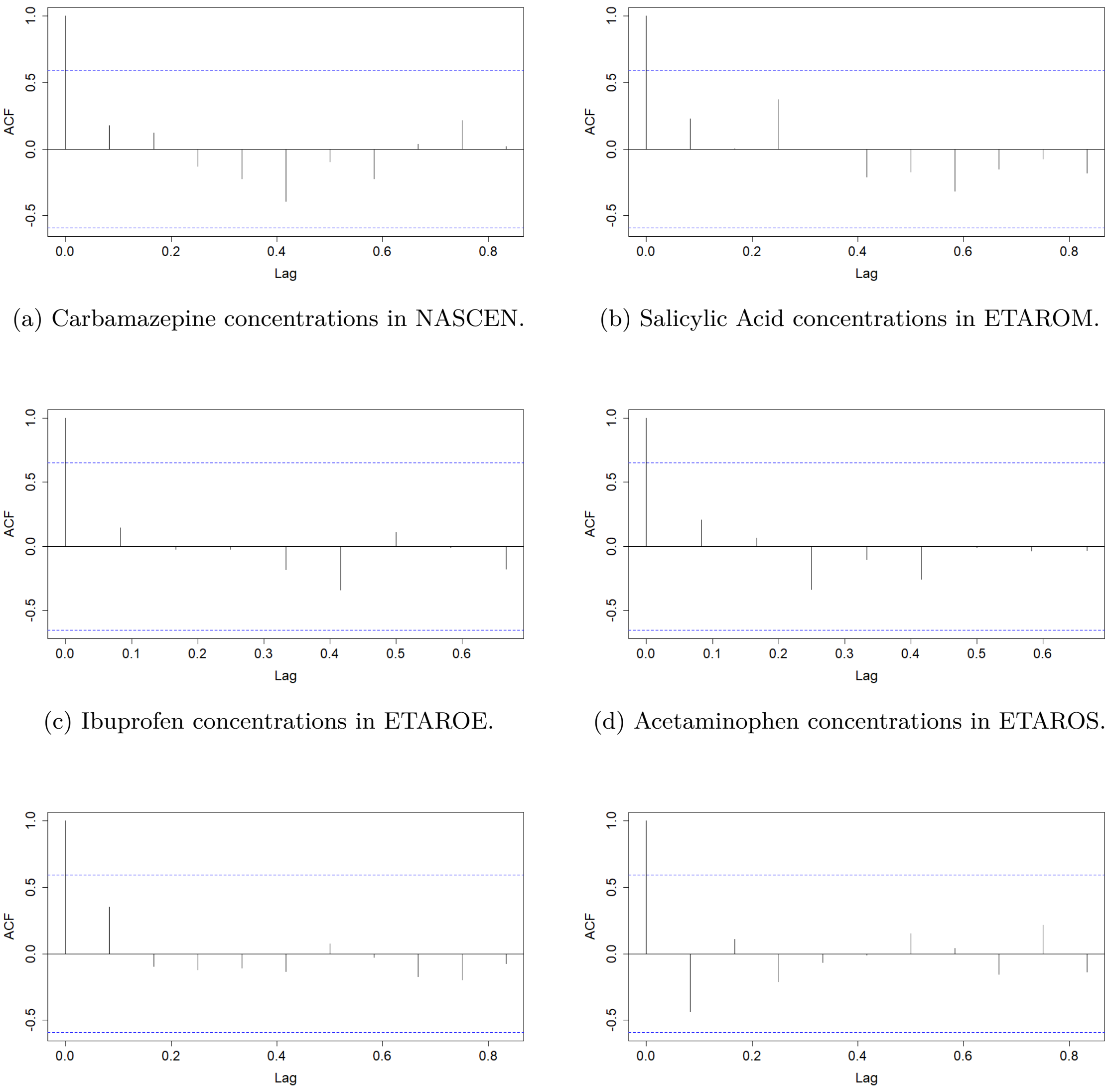
Figure 3.8: Initial sampling dates used for each location.

In fact, changing our data set to a regular time series will bring more opportunities to apply known methods that otherwise would not be possible.

After removing the time series that are always zero or below the detection level, as previously mentioned, we generated Autocorrelation Function (ACF) plots for each pharmaceutical at each sampling site and obtained a total of 155 plots. Figure 3.9 displays the ACF plots obtained for a selected drug in each of the nine sampling points, and that are representative of the overall results.

Unfortunately, as expected, the results of the auto-correlation came to confirm the short time window that we have in all the time series. Almost all the auto-correlation values for the lag do not even pass the visible dashed lines in the plots. This means that almost 100% of the auto-correlation values are not statistically significant, showing no correlation between them.

The low correlation can be justified by the interference of external factors that can have a great impact in the concentrations values found at a given sampling point of the river. One example of such interference is when it rains before the sampling and there is the possibility of more pesticides being dragged from close crop fields.



(a) Carbamazepine concentrations in NASCEN.

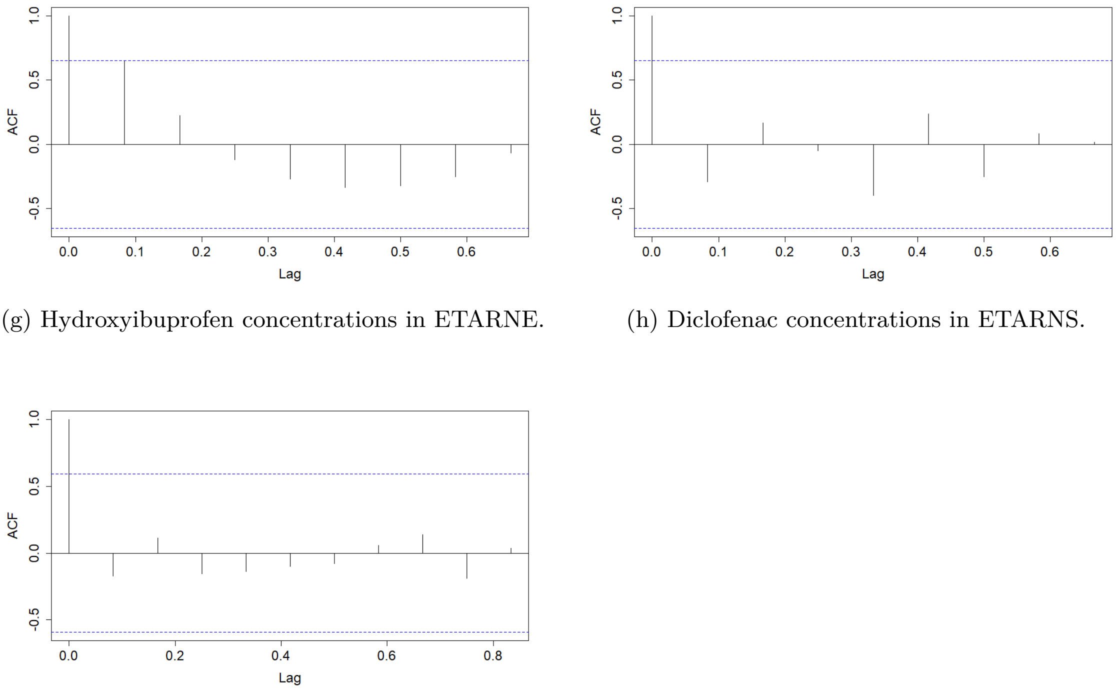
(b) Salicylic Acid concentrations in ETAROM.

(c) Ibuprofen concentrations in ETAROE.

(d) Acetaminophen concentrations in ETAROS.

(e) Naproxen concentrations in ETAROJ.

(f) Fluoxetine concentrations in ETARNM.



(g) Hydroxyibuprofen concentrations in ETARNE.

(h) Diclofenac concentrations in ETARNS.

(i) Venlafaxine concentrations in ETARNJ.

Figure 3.9: ACF plots of some pharmaceuticals in the nine sampling points.

## 3.3 Experiments

### 3.3.1 Experimental Setup

For this first case study we used the train-test partition shown on Table 3.3. Our goal was to have two months for test. Thus, the split percentage was different for two groups of sampling locations, according to their number of observations. The number of observations in the locations on the first group is 9, while on the other group is 11. The split was done trying to keep 80% of the observations for training and 20% of the observations for testing. We then used a sliding-window strategy with last 7 or 9 observations for training, depending on the group of sampling locations, and the following observation for test.

| Sampling Location | Set | Months | Split Percentage |
|---|---|---|---|
| ETAROE \| ETAROS \| ETARNE \| ETARNS | train | October - April | 77.7% |
| | test | May - June | 22.2% |
| NASCEN \| ETAROM \| ETAROJ \| ETARNM \| ETARNJ | train | August - April | 81.81% |
| | test | May - June | 18.18% |

Table 3.3: Split percentage per Sampling Location.

For all the pharmaceuticals time series we started by applying simple methods such as mean forecast, which returns the mean of the time series as prediction, and the random walk method that uses the last observed value as a prediction. Both models use functions of the package `forecast` and serve as baseline for the comparison with more complex models.

We also used the sliding window strategy on more complex forecasting methods, which unlike the previous two, take into account calendar effects, trend and seasonality. We used simple and double exponential smoothing (Holt Method). The Auto-Regression Integrated Moving Average (ARIMA) method was used through the function `auto.arima`, that helped us by automatically choosing the ARIMA parameters. Lastly, a neural network and a Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality (TBATS) model were applied. All the functions used to build these models are available in the package `forecast`.

The choice of methods was heavily influenced by the insufficient amount of available data. Methods such as Holt-Winters were not used given the reduced number of observations we have available.

Using the package Hierachical Time Series (HTS) (Hyndman et al., 2015) we were able to create a hierarchical time series with all the pharmaceuticals time series together with their hierarchy (cf. Figure 3.2. It is important to notice that, the pharmaceuticals that were never detected as mentioned in the previous section, were not considered in for the creation of the HTS model.

One of the first ideas we had, to be able to use in the best possible way the hierarchical structure shown in Figure 3.2, was to predict the future values not for one pharmaceutical but for a whole sub-group. For example, instead of creating predictive models for the Carboxyibuprofen and Hydroxyibuprofen, create instead a prediction model for the sub-group Ibuprofen metabolites, which aggregates both of them. Before we could put this idea into practice, we contacted the specialists in the project that, unfortunately, promptly informed us that although these drugs were organized within common groups, it would not be right to make predictions for one sub-group, with several pharmaceuticals, due to their different behavior.

### 3.3.2 Obtained Results

In Figure 3.10 we have the MAE estimates obtained by each prediction model for each pharmaceutical in a given location. Even though Mean Absolute Percentage Error (MAPE) is one of the most common and preferred measure of forecast error, it was not suitable for this case study. There are some zeros or near-zero values that make impossible to calculate the MAPE (cf. Section 2.3.6.2) In this context, we opted for the Mean Absolute Error (MAE) metric. The box plots shown in Figure 3.10 allow us to understand how the errors are distributed by each prediction model, depending on the location. It is important to mention that not all the errors that were obtained are represented in this graph because some values were of a higher magnitude than the others, making impossible to analyze the rest of the box-plots. The box-plot containing all the errors can be seen in the Appendix A. A brief visual analysis of the graphs let us conclude what has already been said previously that, in fact, there is a great flaw in the quality of the data. The necessary reduction of the y-axis scale, and consequently the removal of data on the plot, was largely due to the presence of fluctuations on the time series, for example huge values followed by near zero values. The model TBATS showed a very poor forecast most likely due to its focus being on time series with multiple frequencies and a higher number of observations.

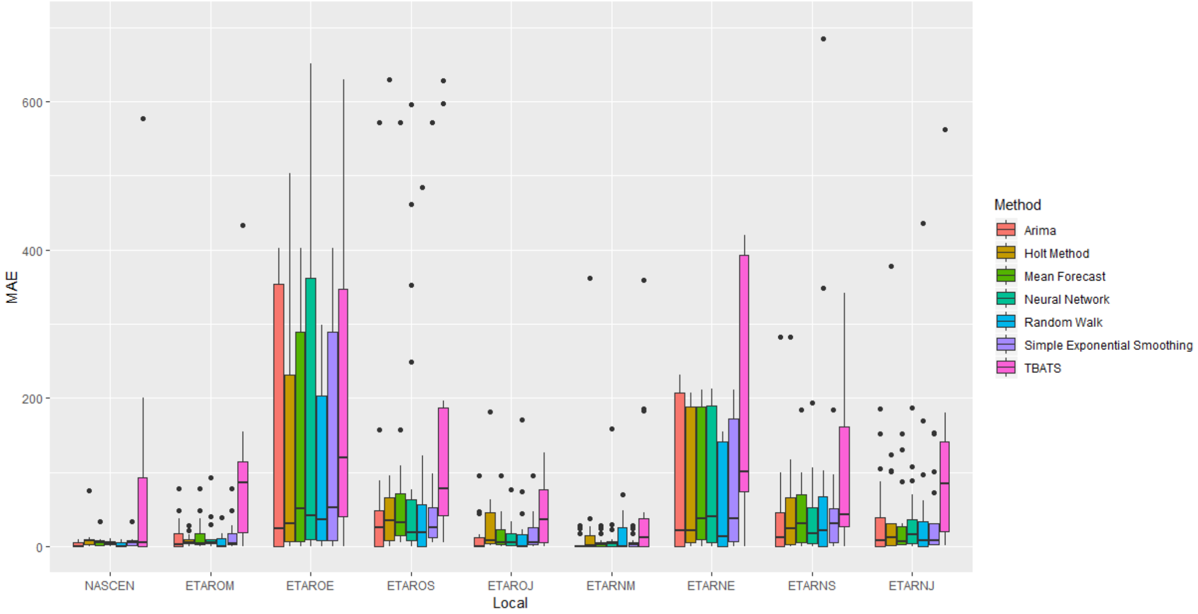


Figure 3.10: Box-plots of MAE estimates obtained by each prediction model for every pharmaceutical in a given location.

Some reasonable error values were obtained but they are mostly due to the small window of training and to the fact that some series are relatively stationary, which makes the prediction of future values easier.

In order to overcome some of the difficulties posed by the insufficient amount of data, we applied the HTS model to forecast the future values for each pharmaceutical and test if, considering the hierarchical structure, the model can get better results than most traditional models. In the Figure 3.11 we can observe the results obtained from the HTS model using Random Walk as the prediction method and the Optimal forecast combination as the method for distributing forecasts within the hierarchy, it was necessary again to not include some outliers in order to make it possible to analyze the box-plots. The other prediction and forecast distribution methods were also tested and included in Appendix A. Still, this was the method that showed the best results and an improvement over previous obtained values without the hierarchy as an input.

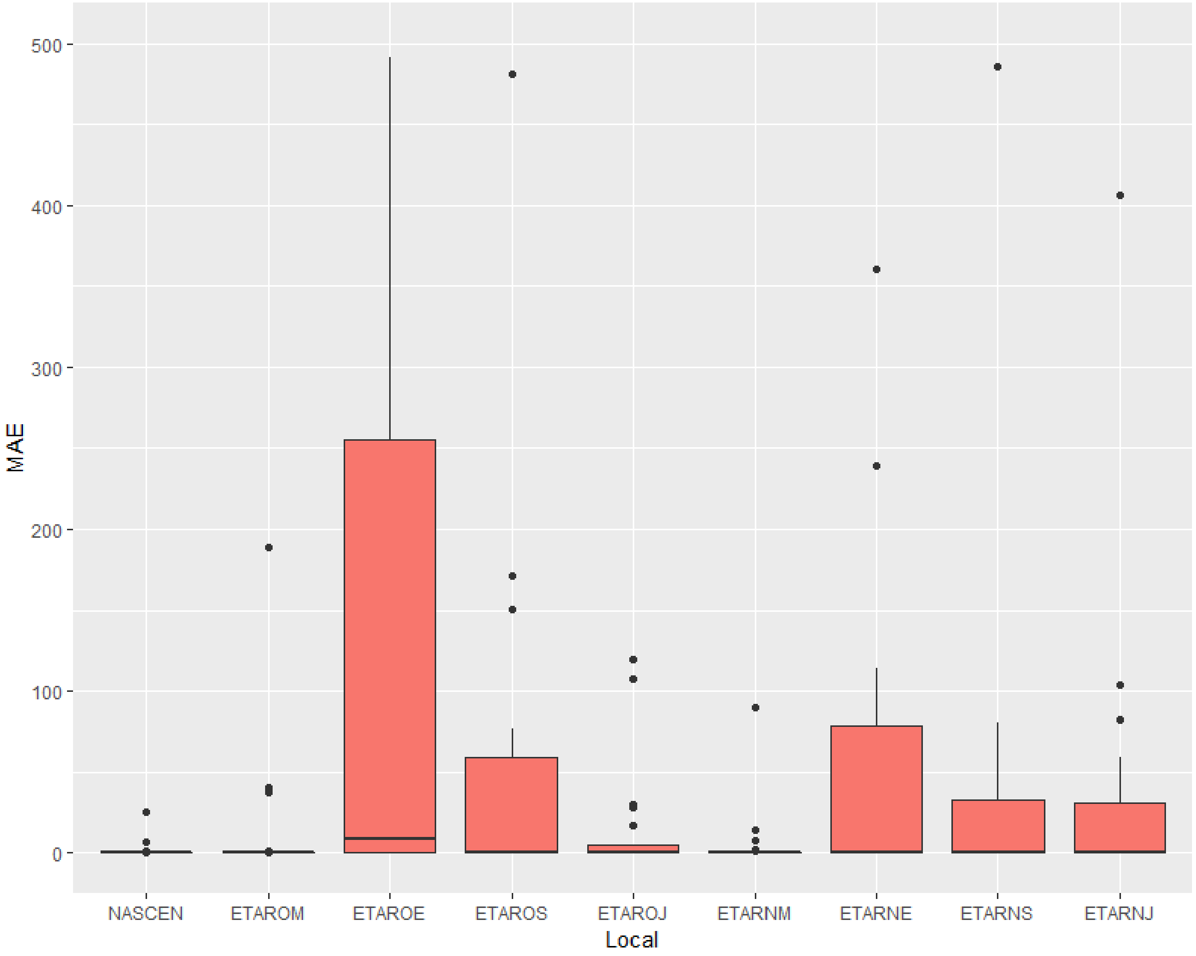


Figure 3.11: Box-plots of MAE estimates obtained by the HTS prediction model for every pharmaceutical in a given location

## 3.4 Discussion

As it became clear in this chapter, the small size of each time series had a negative impact in any of the forecasting models. The number of observations we had for each pharmaceutical was obviously insufficient for any forecasting model to try to understand the behaviour and deduce a trend. Given the small number of training observations, the best that most models can do is to forecast the mean of the past values. Moreover, the observations concerned only one year. Thus, even if we had more observations, no seasonal effects, as we believe that exist in this type of problems, were possible to capture.

Overall, along this case study we came across three principal factors that influenced negatively the obtained results.

First, we were able to obtain zero percentage error while testing some of the models. Unfortunately this is not very accurate, since it happens in time series where all the values are the same. These series exist because there were pharmaceuticals detected always below the method detection limit, and that made us use the values of MDL regarding each pharmaceutical to replace the text “<MDL”.

Then, there was a even bigger problem: the enormous amount of zeros present in our data set. In effect, 69% of the data set was all zeros, a common scenario whenever the data refers to concentration values of pollutants. In these almost zero-constant cases it is expected that the model forecasts zero as it is the constant that minimizes the overall prediction error.

Finally, in addition to the values not being the most helpful to the forecasting task, the amplitude of time they cover is very short, having only 9 months on the WWTP and 11 months on the river, and any sudden change in behavior (even if it is normal as a seasonal effect) in the test window will cause a disastrous forecast and high percentage of error. Even thinking about seasonality is impossible, since it does not even reach one year.

The first two factors would be troublesome even on a big time series. In this case, since we have a very short number of samples, it highlights even more the low quality of the data.

For all these reasons the performance of the tested prediction models was affected and the initial goal of developing prediction models to forecast emerging pollutants in the river could not be reached.

In this context, we searched for other sources of data that could configure a problem similar to this case study but that did not show such limitations. In the following chapter we present a new case study that resulted from this search.

# Chapter 4

# Case Study 2: Forecasting other Pollutants in Water

Our first case study, presented in Chapter 3, revealed to be more troublesome than initially planned. The low number of samples and the range of the concentration values, made it difficult to create a forecast model that was accurate enough to overcome baseline models, like for example, predicting only using the mean of the series. The goal on this chapter is to other data sets, similar to our original data set of pharmaceuticals, that would allowed us to somehow meet the initial goals of this work.

## 4.1 Data set

For the purpose of this case study we resort to two open data sources from which we collected data regarding the concentration values of three other water contaminants usually found in wastewater: Chloride, Caffeine and Lithium.

The first data set is available online and is included in the Heidelberg University's National Center for Water Quality Research (NCWQR) program (Heidelberg University - NCWQR). In the context of this program water samples are taken from several predefined sampling locations, since 1974, with the goal of measuring pollution levels. For our study we used the data regarding Chloride concentration values on the Rock Creek river.

Chloride can be found in the environment as salts of sodium, potassium and calcium. The presence of Chloride in water may be due to human activity, such as industrial effluents, inorganic fertilizers, or road defrosting salts (Evans and Frick, 2001), or from natural causes since it is leached from rocks into soil and water. The threshold for Chloride in drinking water, before it starts interfering with its taste, is in the range of 200-300 mg/litre. Any healthy person will not suffer any harm from intake of large quantity of Chloride provided that there is a simultaneous intake of fresh water (Marlene Evans, 2001).

The portal on water quality of the United States Geological Survey (USGS) (USGS Portal) was the other data source we have explored. The portal offers large data on pollutants and sampling sites in several rivers of the United States of America with records dating back to the year of 1991. Nevertheless, to effectively extract data from this portal requires some domain knowledge to select the appropriate filters. Given the huge amount of data available through USGS portal, exploring and analyzing all the data was not an option. To overcome this obstacle, we decided to use the package `dataRetrieval` (Hirsch and De Cicco, 2015) available on R (Team et al., 2013). This package allows us to retrieve from USGS portal data concerning a given water quality parameter from different sources merged into a single data set. We have chosen data regarding Caffeine and Lithium, each containing time series of the sites where the amount of data was sufficient to be able to carry out an analysis and build prediction models to forecast future concentration values. The data set retrieval was only possible by using the R package `parallel` (Team, 2013), that allowed us to manipulate the number of cpu cores to parallelize the retrieval of the data through the `dataRetrieval` package. The usage of multiple cores to find this data was extremely important due to the size of the USGS database. In Figures 4.1 and 4.2 we can observe the images from Google maps with the location of the sampling points for Caffeine and Lithium, respectively.

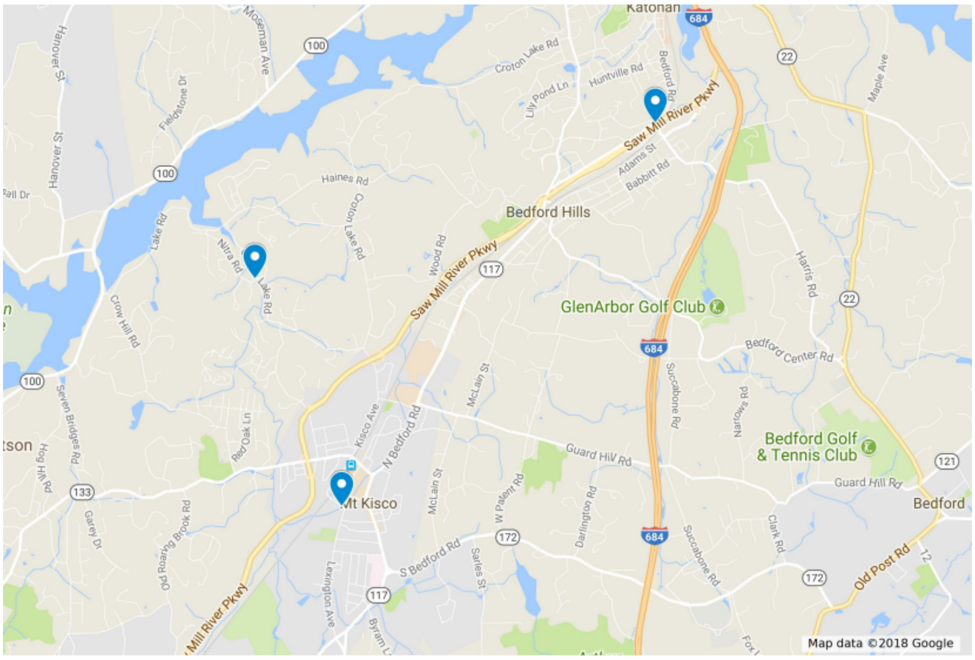


Figure 4.1: Caffeine Sampling Locations

Caffeine is one of the most known and used drugs globally benefiting from a clear social acceptance. The main purpose of this pharmaceutical is to reduce fatigue and give people a wake-up boost. These effects are achieved by stimulating the brain and the nervous system(Burke, 2008). This compound is often used as an indicator of contamination by domestic waste water. The fact is that caffeine is regularly consumed and excreted by the human being, and since it has virtually no origin in the industries, it is considered a human-related contamination (Sauvé et al., 2012, Spence, 2015).

Lithium is a natural compound eluded from rocks and soil into the groundwater. Even though the existing concentrations in nature are lower than those used in therapeutic treatments (Helbich et al., 2015), there are studies that show that the levels of Lithium found in drinking water had an positive impact in reducing suicides rates (Blüml et al., 2013).

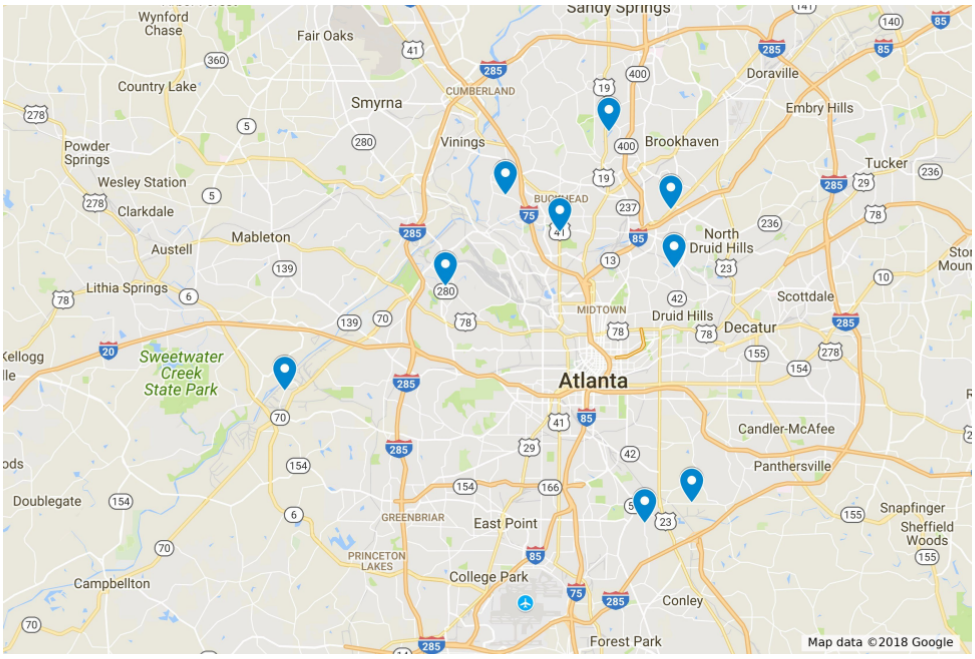


Figure 4.2: Lithium Sampling Locations

## 4.2 Data Analysis and Pre-Processing

The Chloride data set contains daily samples from October of 1982 until October of 2017 (cf. Figure 4.3). These samples correspond to a total of 18910 observations, although it is important to highlight that in some days more than one sample was taken. Since we have a large number of samples, we started by removing the samples concerning the years of 1982 and 2017, so we could have a database with only complete years.To be able to create a regular time series, we transformed our database so that we have daily samples.Thus, in all the days where it was registered more than one sample of the Chloride concentration, only the last sample of the day was maintained. With this change, we had a reduction of 37% in the initial number of samples, having now a remaining 11918 of water samples in total. After having only one sample per day, we observed that not every single day had samples. More precisely, 498 days were missing from our data set. To work around this problem and get a regular time series, we decided to impute values in the missing days by the Last Observation Carried Forward (LOCF) technique available in the R package `imputeTS` (Moritz and Bartz-Beielstein, 2017). Through this technique the missing values are imputed with the value of the last observation.

In the table 4.1 its possible to observe the five-number summary of the data set, i.e. the minimum (Min), the first quartile (Q1), the median, the third quartile (Q3) and the maximum

(Max).

| | **Min** | **Q1** | **Median** | **Q3** | **Max** |
|---|---|---|---|---|---|
| Chloride | 0 | 27.0 | 33.0 | 40.3 | 264.0 |

Table 4.1: Five-number summary of the Chloride data set expressed in mg/L.

Regarding the Caffeine (cf. Figure 4.4) and Lithium (cf. Figure 4.5) data, although they come from the same data source and have an equal structure, they have different sample locations. Unlike Chloride data, no pre-processing is required to be able to create a regular time series, as the observation period is equally separated in months. The remaining Caffeine and Lithium time series can be observed in the Appendix B.

The Caffeine and Lithium data have 3 and 9 time series, respectively. Each series is from a different sample location. The data was transformed so that all series related to the same variable had the same sampling period, thus allowing a more easily a comparison between them. With this change, the data had a total of 80 samples for Caffeine starting in July of 2001 until February of 2008 and a total of 77 samples of Lithium starting in August of 2003 and ending in December of 2009. A summary of this two data sets can be found in the tables 4.2 and 4.3.

| **Time Series** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
|---|---|---|---|---|---|
| Caffeine 1 | 0.008 | 0.046 | 0.061 | 0.090 | 0.272 |
| Caffeine 2 | 0.021 | 0.083 | 0.194 | 0.566 | 8.110 |
| Caffeine 3 | 0.022 | 0.068 | 0.117 | 0.311 | 4.600 |

Table 4.2: Five-number summary of the Caffeine data set expressed in $\mu$g/L.

| **Time Series** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
|---|---|---|---|---|---|
| Lithium 1 | 6.000 | 29.500 | 36.875 | 42.000 | 109.800 |
| Lithium 2 | 14.333 | 23.000 | 26.500 | 30.333 | 59.000 |
| Lithium 3 | 13.000 | 28.560 | 32.500 | 34.750 | 51.000 |
| Lithium 4 | 6.000 | 24.952 | 29.000 | 36.500 | 48.000 |
| Lithium 5 | 15.000 | 24.500 | 28.500 | 34.400 | 41.000 |
| Lithium 6 | 16.000 | 24.320 | 29.875 | 34.667 | 45.000 |
| Lithium 7 | 17.071 | 27.667 | 31.500 | 34.472 | 41.000 |
| Lithium 8 | 21.000 | 34.400 | 43.000 | 49.667 | 63.000 |
| Lithium 9 | 14.000 | 32.888 | 36.222 | 39.250 | 47.500 |

Table 4.3: Five-number summary of the Lithium data set expressed in $\mu$g/L.

## 4.3 Experiments

### 4.3.1 Experimental Setup

In this second case study, Monte Carlo simulation was chosen by us as a methodology to try to compare the different prediction models across the largest window of data possible. For this, we selected 10 random dates for each one of the data sets, since all three have different starts and endings, and the Chloride data set consists in daily data against the monthly data from Caffeine and Lithium. For all the randomly generated dates, we used the last two years of data for training and the next 6 months to test the prediction model, all the dates generated for the three different data sets can be seen in figs. 4.3 to 4.5.

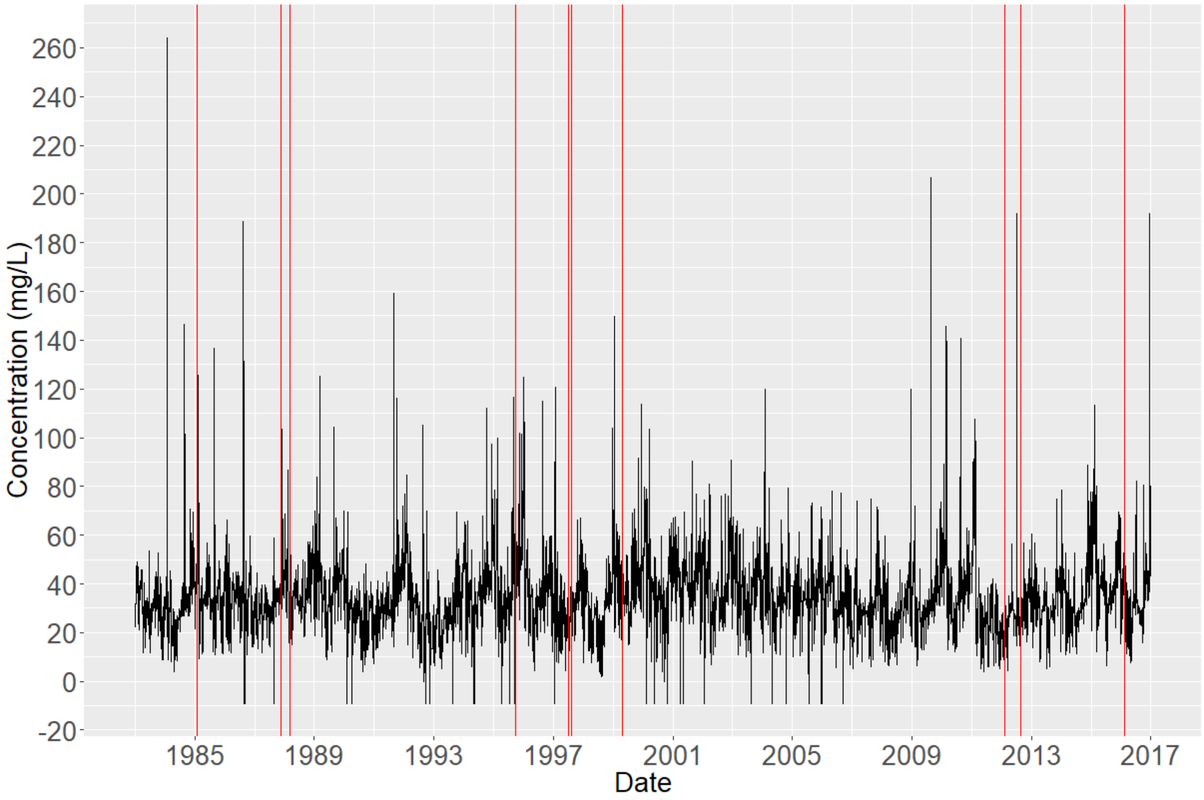


Figure 4.3: The 10 random dates used by the monte carlo simulation on Chloride time series..

A sliding window strategy was used for all the forecasts, from time series and embedded series. Important to note that, for the monthly data (Caffeine and Lithium), the models were re-constructed every month using the past $N$ months to create the model, where $N$ is the size of the train set. For the Chloride daily data, the model was re-constructed every 30 days using the past 30 days to construct the next model.

To enable the application of regression methods, we also transformed the time series into embedded time series. This consisted in mapping each time series into embedding vectors

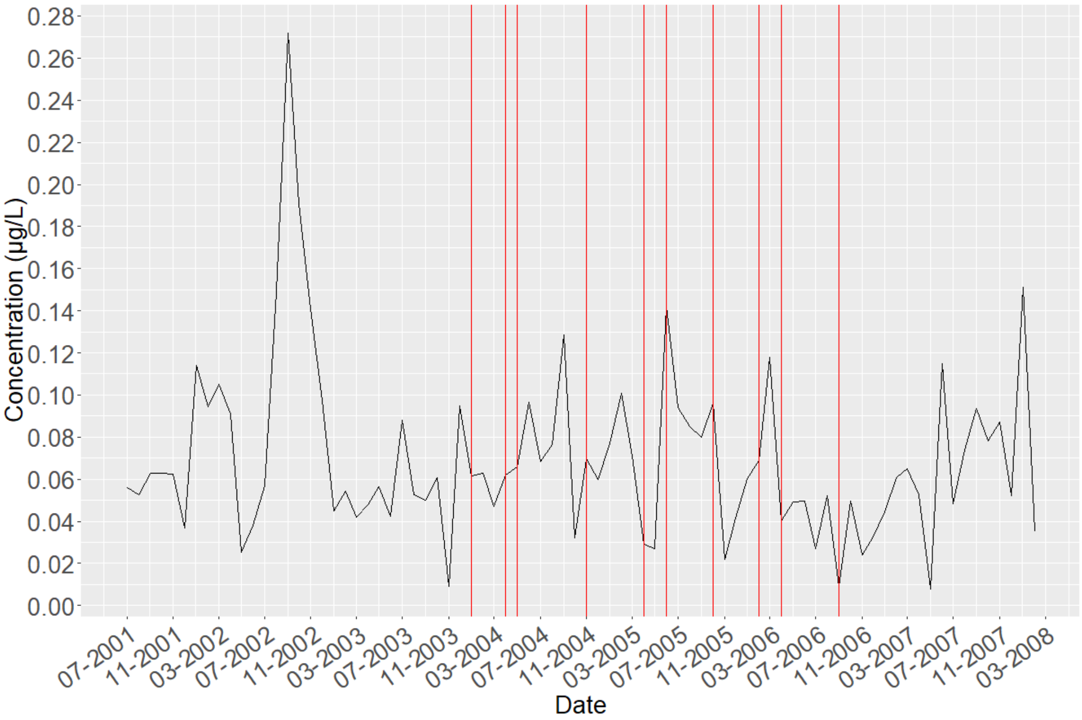


Figure 4.4: The 10 random dates used by the monte carlo simulation on Caffeine 1 time series.

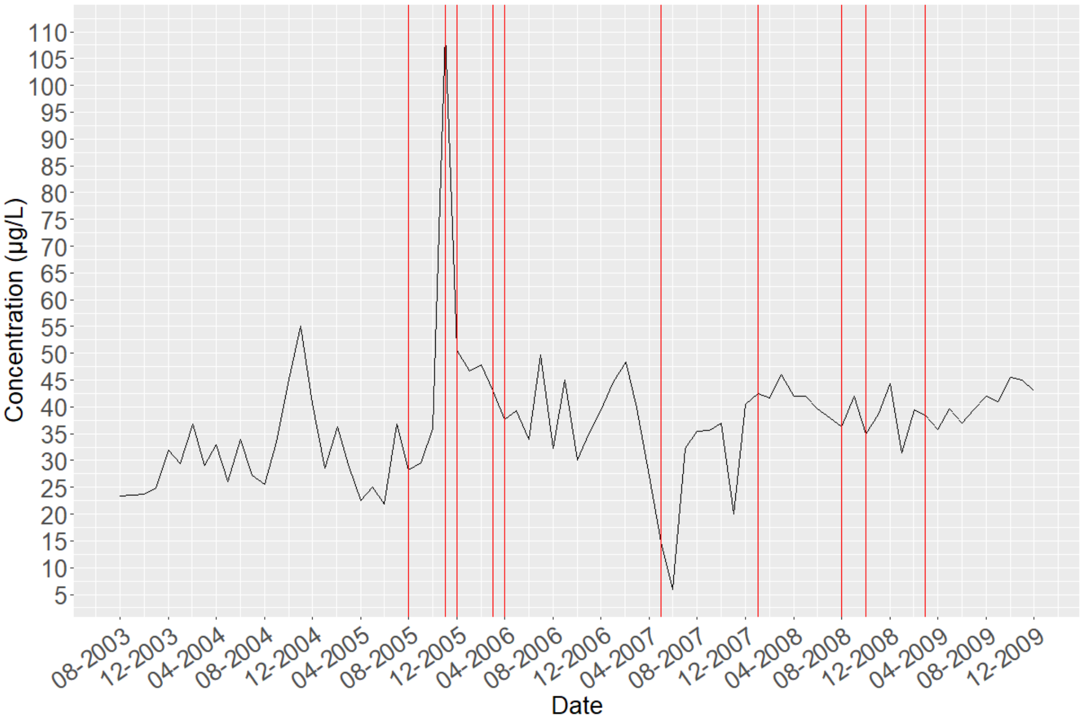


Figure 4.5: The 10 random dates used by the monte carlo simulation on Lithium 1 time series.

containing the observation and the $k$ past observations. The time delay embedding series were created using the last 30 days for the daily data on Chloride and the last 6 months on the Caffeine and Lithium monthly data.

For this second case study we applied the same methods we used in the first case study and also the Holt-Winters method. The number of available observations makes now possible to use this method. Regarding the regression methods, we tested Arbitrated Dynamic Ensemble (ADE) ensemble approach from the package `tsensembler` (Cerqueira et al., 2017), Support Vector machines from `e1071` package (Meyer and Wien, 2001), Random Forest (Liaw et al., 2002) and RPART (Recursive Partitioning and Regression Trees) (Therneau et al., 2010), from the packages `randomForest` and `rpart`, respectively.

Despite the reduced information available on these data sets, one thing we noticed by analyzing the coordinates of each Lithium sampling location is that, at least three of them, in a total of nine, might have spatial-temporal correlation between them. This can be observed in the Figure 4.6. The green and orange marker point to the North Fork Peachtree Creek river and the South Fork Peachtree Creek, respectively. The two rivers join together in the Peachtree Creek river where our third sampling point is located with the blue marker. Thus, in our experiments we also explored for this data sets the impact of including the spatial information in the performance of the forecasting models.

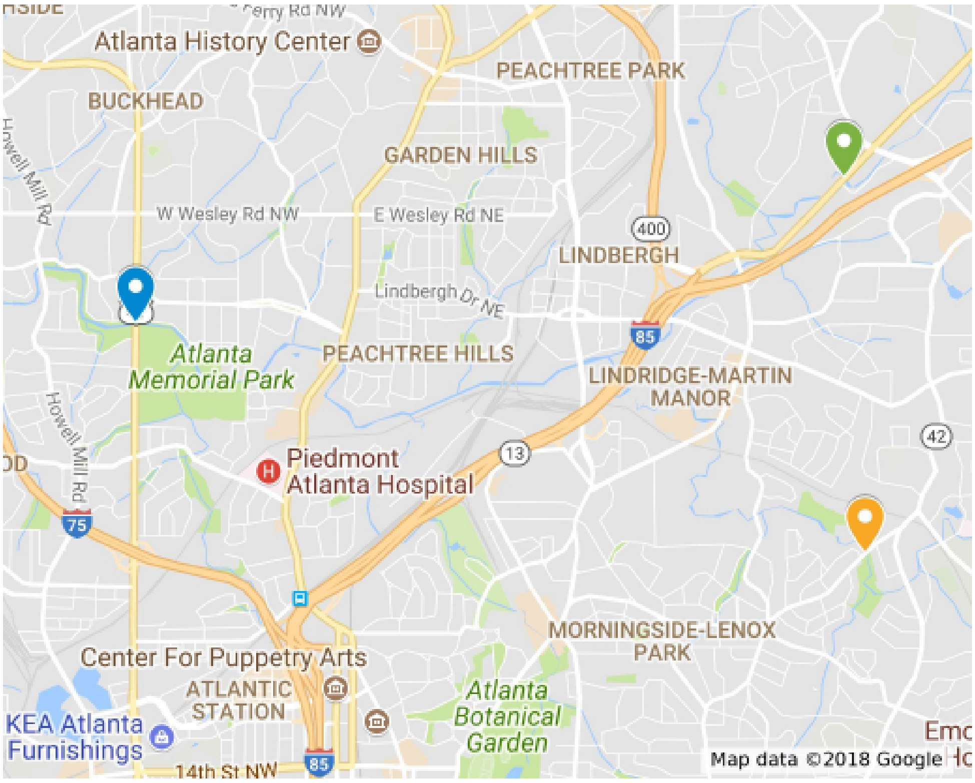


Figure 4.6: Map of three Lithium sampling locations with spatial-temporal correlation.

### 4.3.2 Obtained Results

We start by analyzing the results of applying the sliding window method in conjunction with monte carlo simulation on the time series. Figure 4.7 shows the box-plots of Mean Absolute Error (MAE) estimates obtained by each prediction method for the Chloride data set.

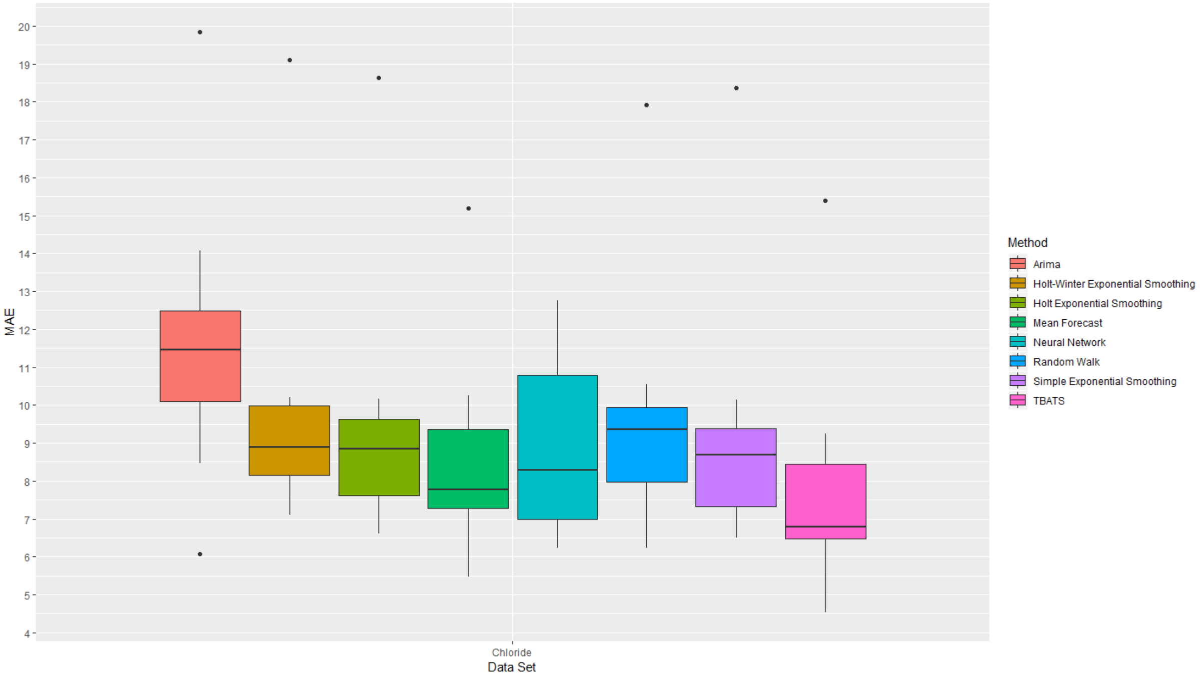


Figure 4.7: Box-plots of MAE estimates obtained by the prediction models for Chloride

In the Figures 4.8 and 4.9 we have also the box-plots of MAE estimates obtained by each prediction method but, this time, for the Caffeine data set. We had to separate this data set in two different plots because, even though the sampling locations are close to each other (see Figure 4.1), the time series of the "Caffeine 1" has considerably lower values.

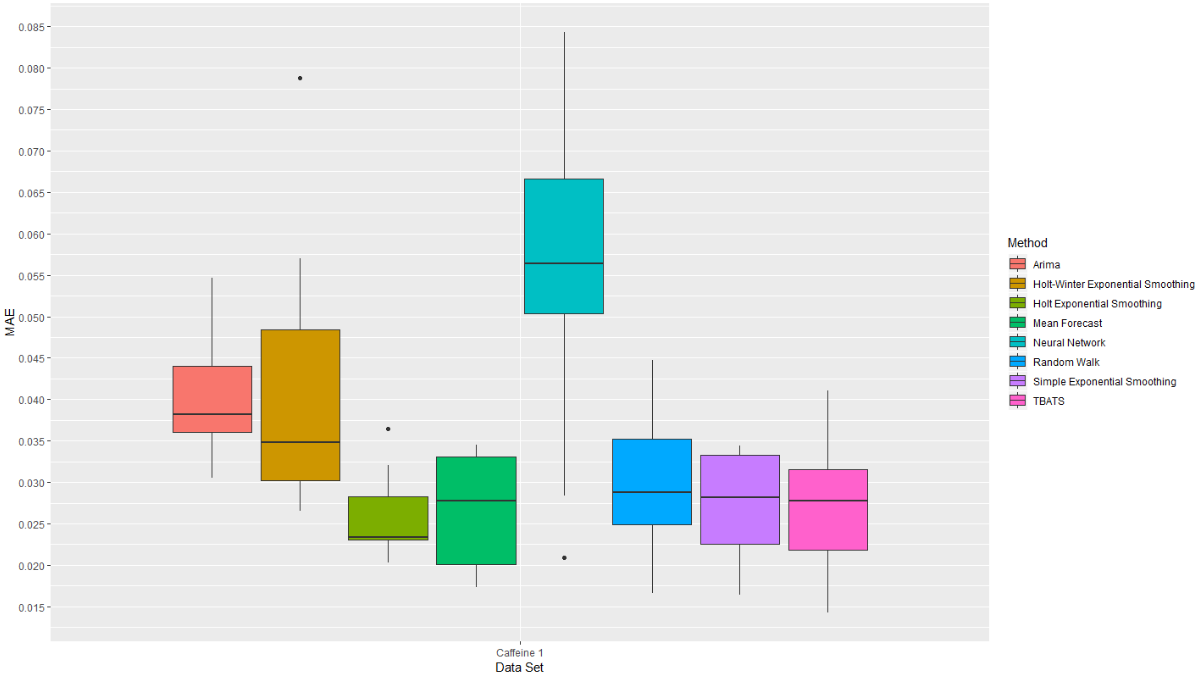


Figure 4.8: Box-plots for MAE estimates obtained by the prediction models for Caffeine 1

Lastly, we have in the Figure 4.10 the box-plots of MAE estimates obtained by each prediction

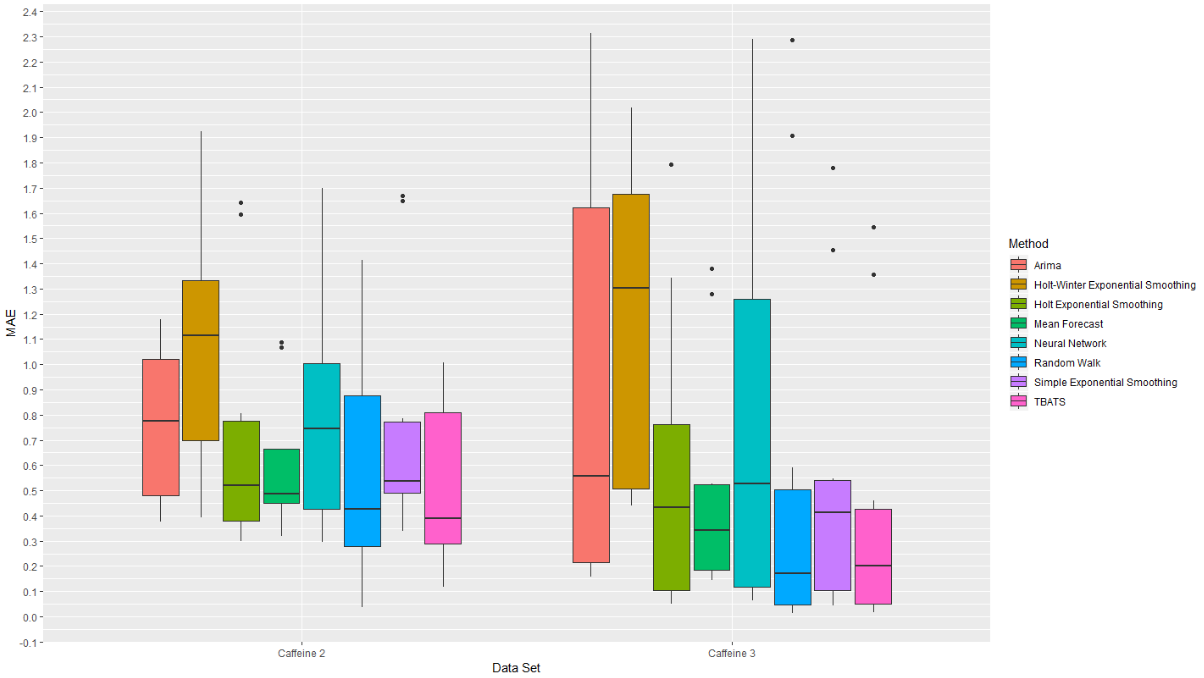


Figure 4.9: Box-plots of MAE estimates obtained by the prediction models for Caffeine 2 and Caffeine 3

model for all the nine sampling locations for the contaminant Lithium.

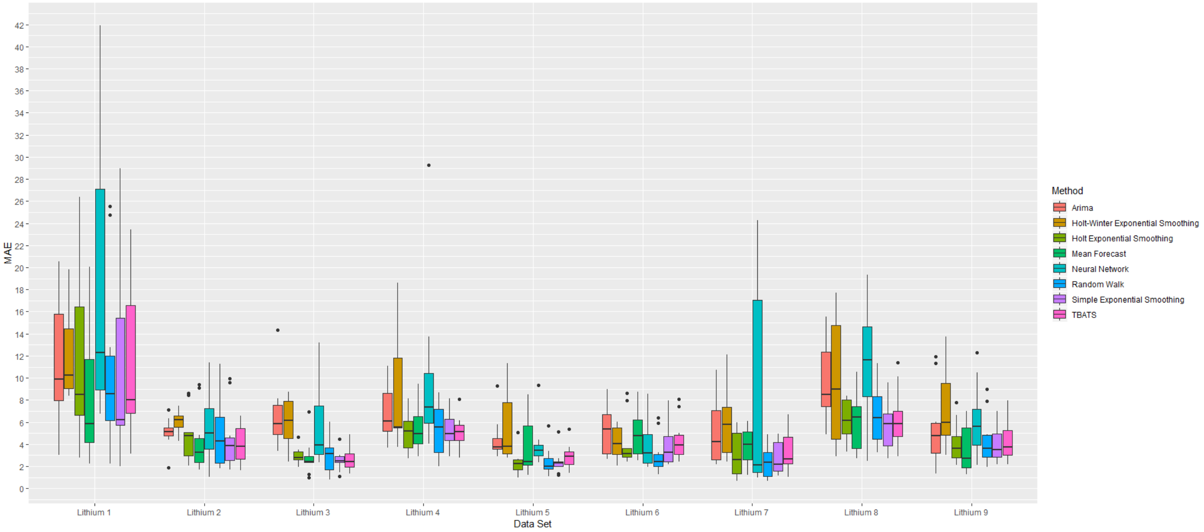


Figure 4.10: Box-plots of MAE estimates obtained by the prediction models for all the time series in the Lithium data set

It is important to highlight that all these error estimates are dependent on the magnitude of the data, because we used the MAE. Therefore, to analyze the error measure it is necessary to look at the tables 4.1 to 4.3, in order to better understand the quality of the prediction method.

Overall, the results obtained using the Auto-Regression Integrated Moving Average (ARIMA) model were disappointing across all data sets. Even though in some time series the results were closer to the rest, like for example in Caffeine 2, it was always one of the worst. The Chloride data set is the most significant example of poor results from this forecasting method.

The prediction method Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality (TBATS), a fairly new model specially when compared to more traditional ones like ARIMA or Random Forest, showed good results in all time series, as is easily observable in the Chloride data set (cf. Figure 4.7). This came with no surprise, since the reason behind the creation of TBATS model was to deal with multiple seasonal periods. Therefore, our daily data on the contaminant Chloride was a perfect data set to show how the model can help us improve and better predict the future values.

Regarding the results obtained for the Lithium data set, we can see that there is a large increase in the error value for the time series Lithium 1 and a slight increase for the Lithium 4 and Lithium 8. The large error for the first time series can be explained by the existence of outliers (see Figure 4.5). As has been said previously, the monte carlo experiment allows us to test the prediction models in a bigger percentage of the data set, instead of only picking one training set and on test set from one date. Due to the random dates, there is the chance that the outlier or outliers will be present more than one time in either the training set or the data set, depending on the date that was chosen. Undoubtedly, the presence of outliers in the train or in the test data affects either the construction of the prediction model or the error estimates. Still, and specially in this case, we cannot completely ignore their presence since it can have a huge impact in human wealth.

As for the time series Caffeine 4 and Caffeine 8, their slightly worst results when compared to the rest of the time series in the Lithium data set, can be associated to the more distinct and strong trend that is present in these two time series. Using the Seasonal and Trend decomposition using Loess (STL) (Cleveland et al., 1990) of R, we were able to extract the trend for both of the time series. By analyzing Figure 4.11, we can better understand the significant growth in both of these series.

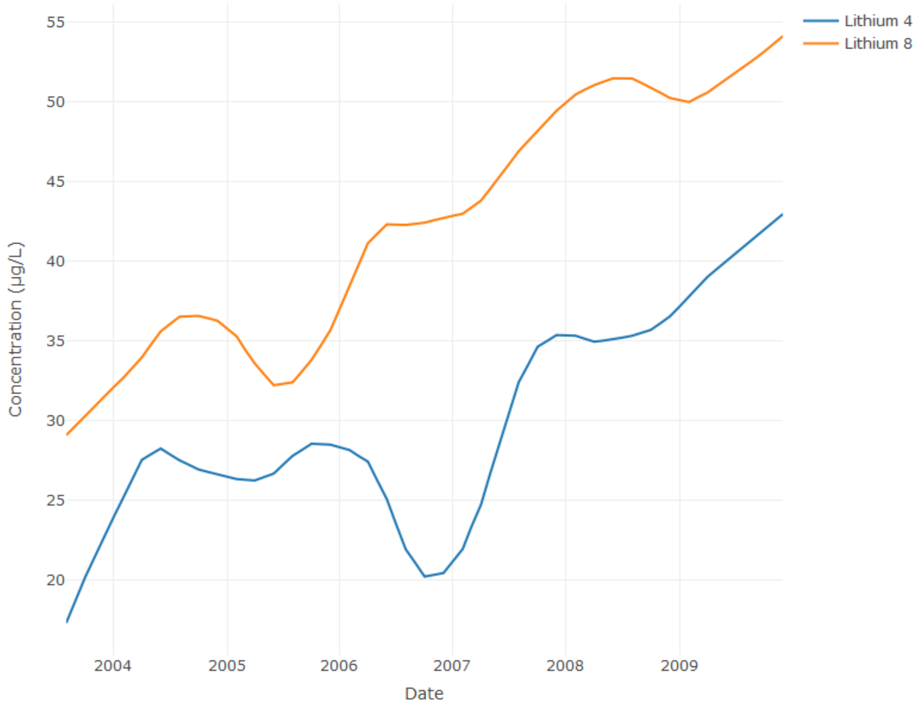


Figure 4.11: Trend component of the Lithium 4 and Lithium 8 time series

The results obtained for the embedded series can be seen in Table 4.4. In general, the results were more accurate in the embedded series, although some methods, such as the TBATS, were able to obtain better results in some of the time series when compared to the results of the regression methods.

Unfortunately, the Lithium 1 time series continues to demonstrate high errors due to the difficulties in modeling the outliers, as it can be seen by the MAE estimates obtained by all the four regression models.

On the positive side, the time series of the contaminant Chloride showed great improvement (cf. Figure 4.12), reducing almost in half the MAE estimate given by the regression methods when compared to the ones initially applied to the time series. The reason why this Chloride time series stands out when we use the embedded series is probably due to the fact that regression methods work a lot better if they have a large training window. The rest of the MAE box-plots for this case study are in Appendix B.

| | Random Forest | RPART | SVM | Tsensembler |
|---|---|---|---|---|
| Chloride | 4.5333 | 4.9343 | 4.9064 | 4.3368 |
| Caffeine 1 | 0.0286 | 0.0308 | 0.0298 | 0.0287 |
| Caffeine 2 | 0.5543 | 0.4747 | 0.3161 | 0.4358 |
| Caffeine 3 | 0.4675 | 0.4968 | 0.3658 | 0.4999 |
| Lithium 1 | 10.3778 | 10.2442 | 7.6043 | 12.1030 |
| Lithium 2 | 4.5379 | 5.2871 | 3.2743 | 4.8658 |
| Lithium 3 | 5.0003 | 5.3742 | 4.6349 | 5.0415 |
| Lithium 4 | 6.8897 | 6.2835 | 6.8212 | 7.2166 |
| Lithium 5 | 3.5988 | 4.2315 | 3.7525 | 4.4656 |
| Lithium 6 | 3.6647 | 3.8469 | 3.9932 | 4.2316 |
| Lithium 7 | 2.5348 | 2.3260 | 2.3952 | 2.6545 |
| Lithium 8 | 5.6529 | 5.8873 | 5.9336 | 6.2654 |
| Lithium 9 | 6.2768 | 5.6576 | 6.1322 | 7.1279 |

Table 4.4: MAE estimates obtained by each prediction model on all the embedded series.

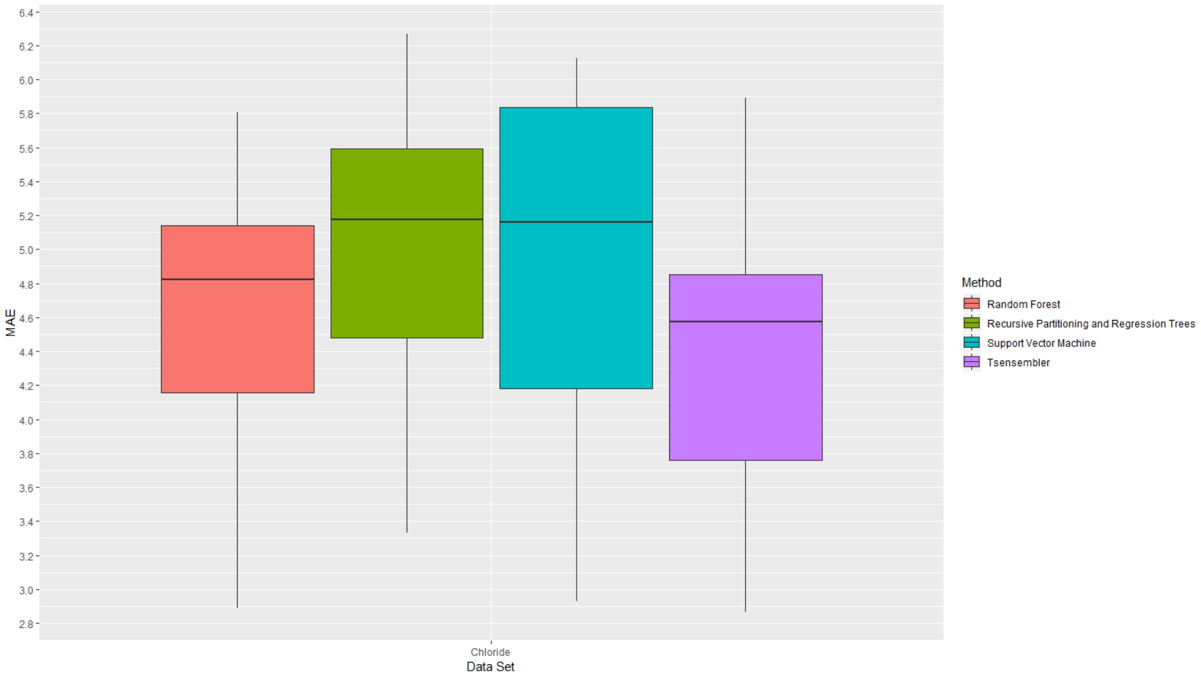


Figure 4.12: Box-plots of MAE estimates obtained by the prediction models for the embedded Chloride data set

The embedded series also allowed us to explore the spatial component of the data set in order to try to improve the forecast. More specifically, we are referring to the time series Lithium 6 signaled by a blue marker in Figure 4.6. In order to test the impact of including the spatial component, we used three different types of embedded time series, as follows.

**Original Series** - The embedded time series of Lithium 6, that was our target to forecast.

**One Spatial Component** - An embedded time series using the current value and the last two observations of Lithium 6, together with two past observations of another time series (Lithium 5 or Lithium 7).

**Two Spatial Components** - An embedded time series using the current value and the last two observations of Lithium 6, together with two past observations from the two time series (Lithium 5 and Lithium 7).

In Figure 4.13 we have the MAE results of the forecasts for all the above described setups.

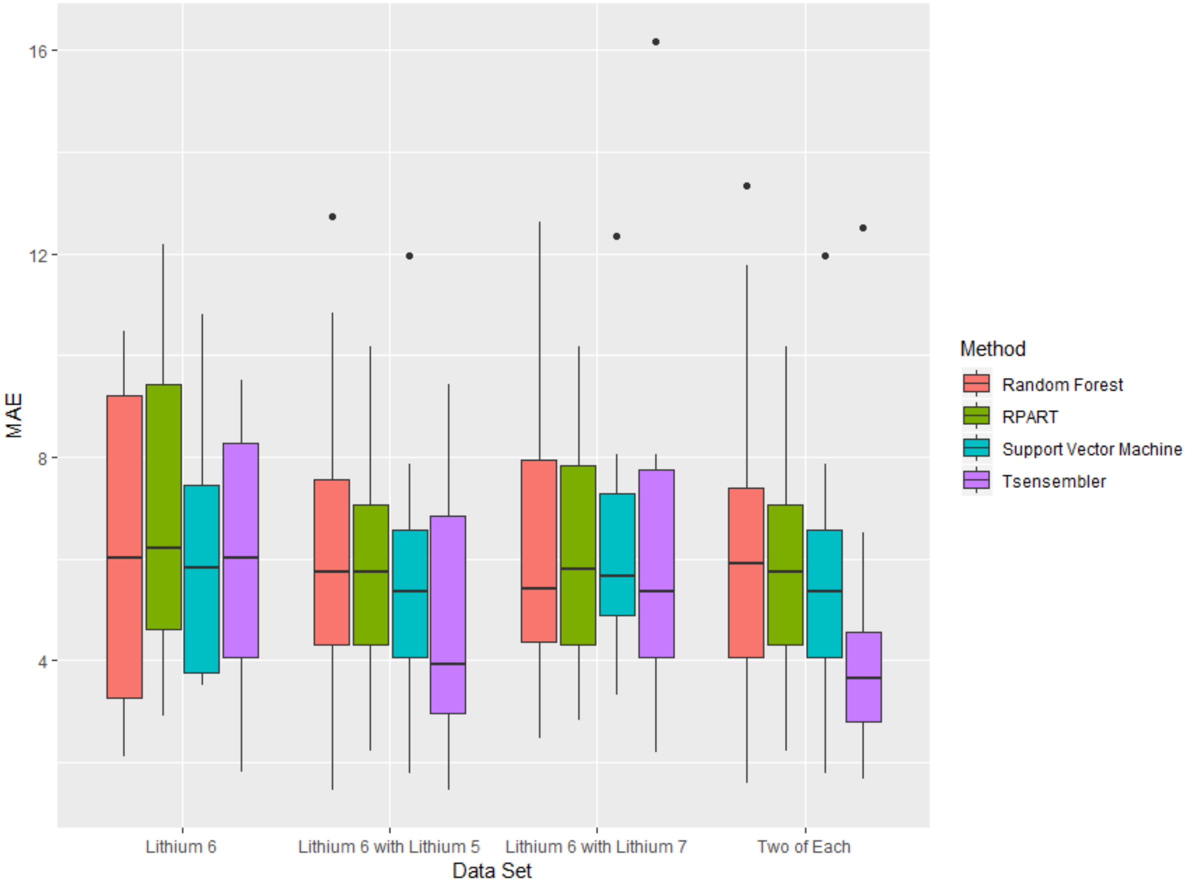


Figure 4.13: Box-plots of MAE estimates obtained by the prediction models in four tested embedded series.

As we can see, we achieved an improvement on the forecasts accuracy by using a equal number of past observations from the target time series and the other two in the embedded series. It is important to highlight that the Lithium 6 time series together with Lithium 5 time series (embedded series using two past observations of each one) gave good results when compared to only use the Lithium 6 embedded series. This can be explained by the close trend behaviour of both time series, as it is shown in Figure 4.14. Moreover, this also justifies why the setup that includes only the Lithium 7 achieves slightly worse results than the one that includes only the Lithium 5.

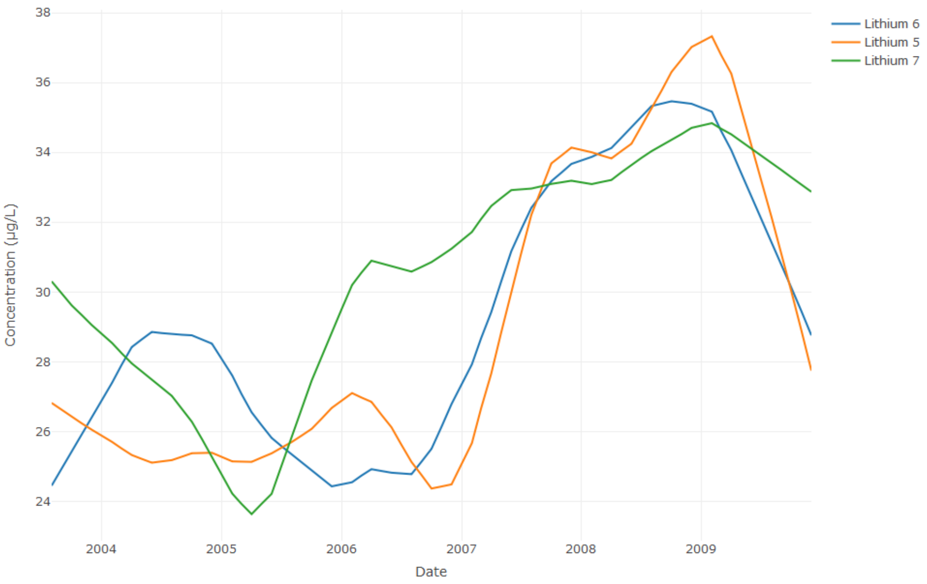


Figure 4.14: Trend component of the Lithium 5, Lithium 6 and Lithium 7 time series

## 4.4 Discussion

The initial goal of this case study was to find data sets that shared similar characteristics to the one in the first case study. Unfortunately, we could not find any free access data sets on pharmaceuticals involving Wastewater Treatment Plants (WWTPs) and rivers samples. Instead, we found data sets from several rivers in the United States of America, with concentration values on other pollutants. These data sets offered samples taken with higher frequency and over a larger period of time. This large number of observations allowed us to add seasonality and trend component to the forecasting models.

The results obtained in this case study for both, the time series and embedded time series, were a nice surprise when compared to the first case study. In effect, the apparent good results obtained in the last case study had mainly to do with the almost constant series and, thus, to the overfitting of the forecasting models. The data used in this case study did not have that almost constant character and, still, we were able to achieve good accuracy out of the prediction models.

In order to take advantage of all the information we had about the data, we feed the forecasting models with the spatial information we were able to identify from the map of the sampling locations. This turned out to be a successful model that showed the importance that the two rivers that come together in one, can have in the forecasting of concentration values in that river. Given this indicative results, other time lags can now be further tested to improve the performance of the model. Still, the information of the best time lag to use in the spatial and temporal component is very domain specific. As such, some guidance should be obtained from the domain experts for this process.

# Chapter 5

# Conclusions

Although pharmaceuticals have been present in our drinking water for decades only now, with the advances in technology, their levels in the environment started to be detected and quantified as a potential danger to the human health. Their presence, even at low concentrations, is raising concerns about the effects they have in our drinking water quality and on the aquatic environments. Since they are still unregulated and most of the Wastewater Treatment Plants (WWTPs) are not prepared to treat them, they are considered as emerging pollutants.

The initial goal of this thesis was to create a data mining system that, by receiving as input samples collected at the Lis River, Portugal, would monitor and forecast the presence of pharmaceuticals in the water. Ultimately, the idea is that extreme values of concentration of such pollutants are anticipated so that both authorities and WWTPs can take preventive actions in order to avoid deterioration in the drinking water quality and subsequent impacts in the human health.

We tried different approaches to build a forecasting model that would work the best with the available data. Still, the data was insufficient to be able to obtain good results.The models gave us "false" good forecasts, because the model in some time series would just keep overfitting the training data due to the amount of repeated or zero values. Another important fact was the existing fluctuations in the data, possibly due to external factors. The number observations available for training was so small that the model could not learn this fluctuations, yielding to very poor results.

We also attempted to take advantage of the hierarchical structure of the pharmaceuticals and incorporate it in the forecasting model. Although the above mentioned limitations imposed by the data set, we could see a slightly improvement in the forecast values.

In order to still try to reach our goal, we created a second case study. In this one, we found some free data sources that allowed us to retrieve new data sets with similar characteristics to the Pharmaceuticals data set. With these data sets we were able to try some new approaches in addition to the ones already used in the first case study, since the limitations imposed by the first data set were no longer present. In particular, we were able to obtain good results by the

TBATS algorithm, specially in the Chloride daily data, making use of being capable of changing seasonality slowly over time. On the embedded time series, we saw the Tsensembler model achieved good results by combining several prediction models into one.

## 5.1 Limitations and Future Work

The results obtained in the course of this thesis, specially in the case study 2, give us confidence for a strong future work, even though the main goals were not completely reached, because of several data set limitations.

The insufficient data on the pharmaceuticals was a big barrier for us when building our prediction models. A future plan could be continuing the sampling in Lis river, so that, we could build a bigger data set of pharmaceuticals concentrations. This would allow us to study the seasonality and trend on these emerging pollutants, something impossible at the moment. In a future work, we could also improve the equipment behind the analysis of the samples since, as indicated by one specialist in the project, the new equipment models have a much a lower Method Detection Limit (MDL). That would allow to instead of replacing the "<MDL" for a fixed value, to actually use the real concentration value.

Another goal of the project was to specifically anticipate high extreme values (outliers) of concentrations of those pharmaceuticals pollutants, and not just forecasting the near future. This was not possible to us in the first case study, due to data restrictions we had available, and for the second case, since it was at a later stage of this thesis, we did not have enough time to build an efficient and reliable outlier prediction model.

# Appendix A

# Case Study 1

In this Appendix we have the following graph that is the Mean Absolute Error (MAE) estimates for the case study 1 forecast including all the outliers.

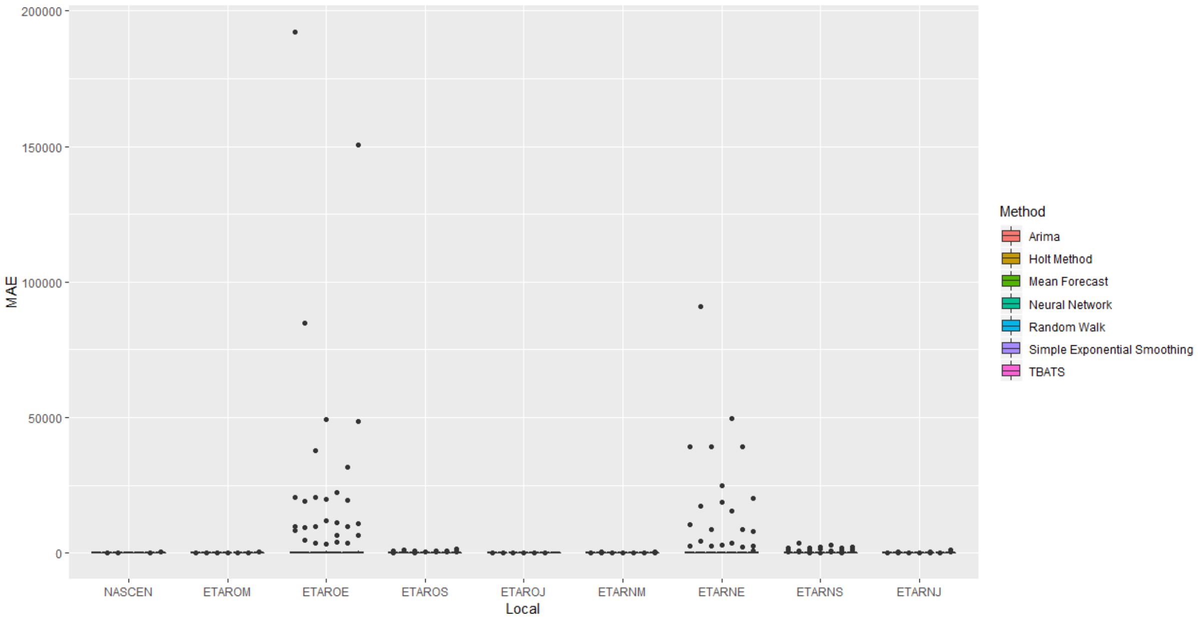


Figure A.1: Box-plots of MAE estimates obtained by each prediction model for every pharmaceutical in a given location.

The following box-plots are the remaining prediction and forecast distribution methods also used in the Hierachical Time Series (HTS) model.

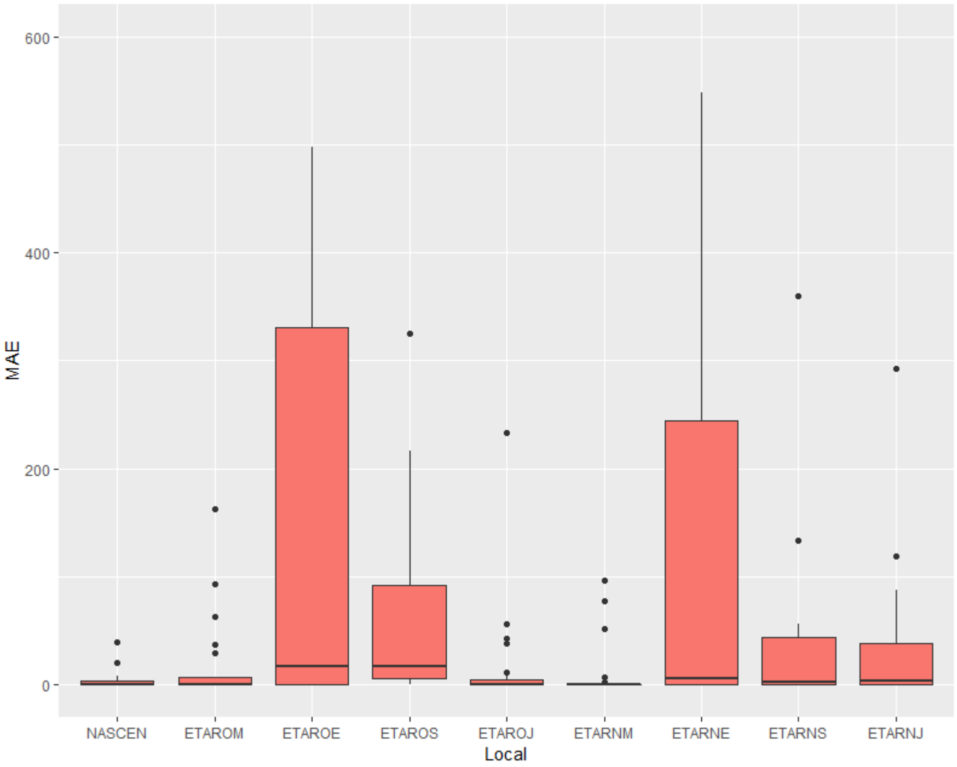


Figure A.2: Box-plots of the MAE estimates obtained using Arima and Optimal forecast combination.

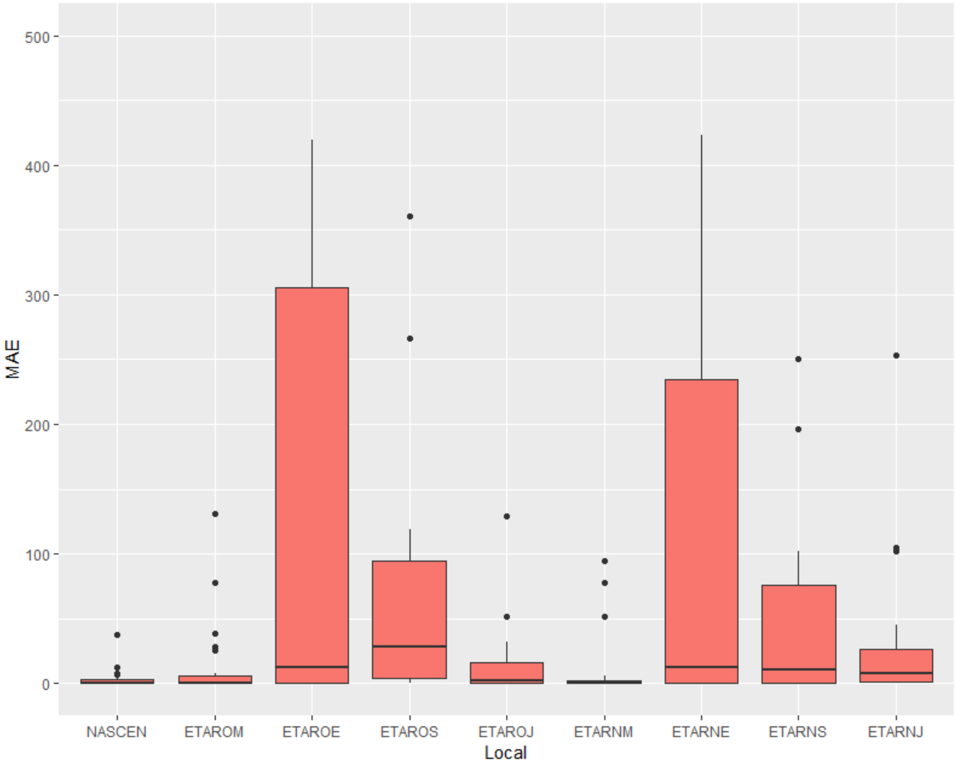


Figure A.3: Box-plots of the MAE estimates obtained using ETS and Optimal forecast combination.

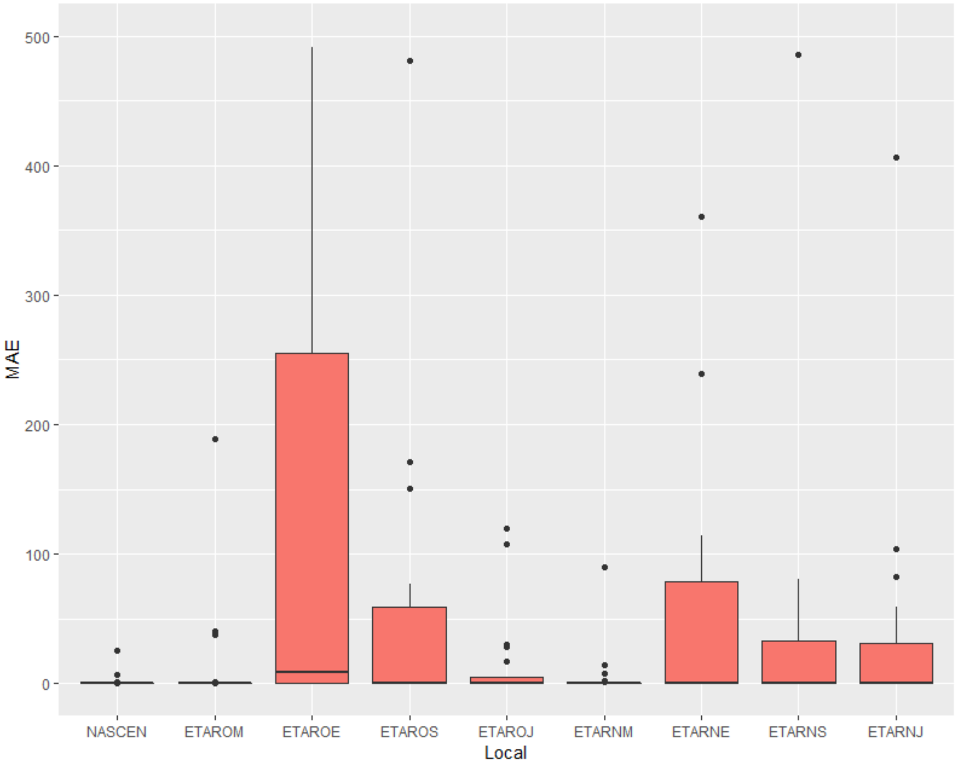


Figure A.4: Box-plots of the MAE estimates obtained using Random Walk and Top-down approach.

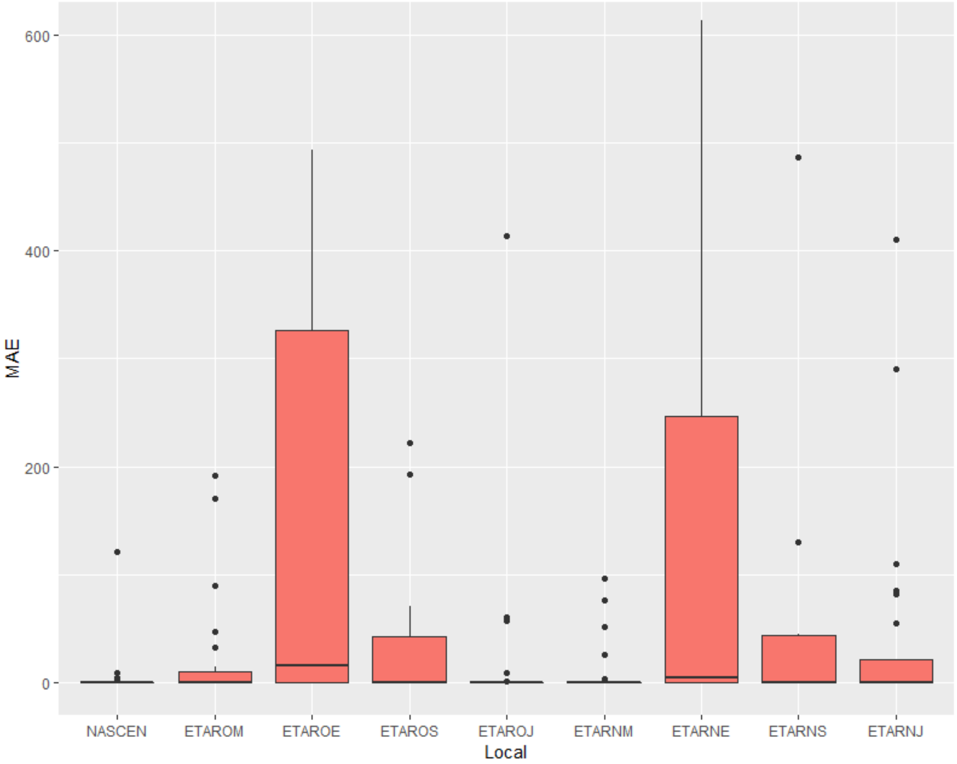


Figure A.5: Box-plots of the MAE estimates obtained using Arima and Top-down approach.

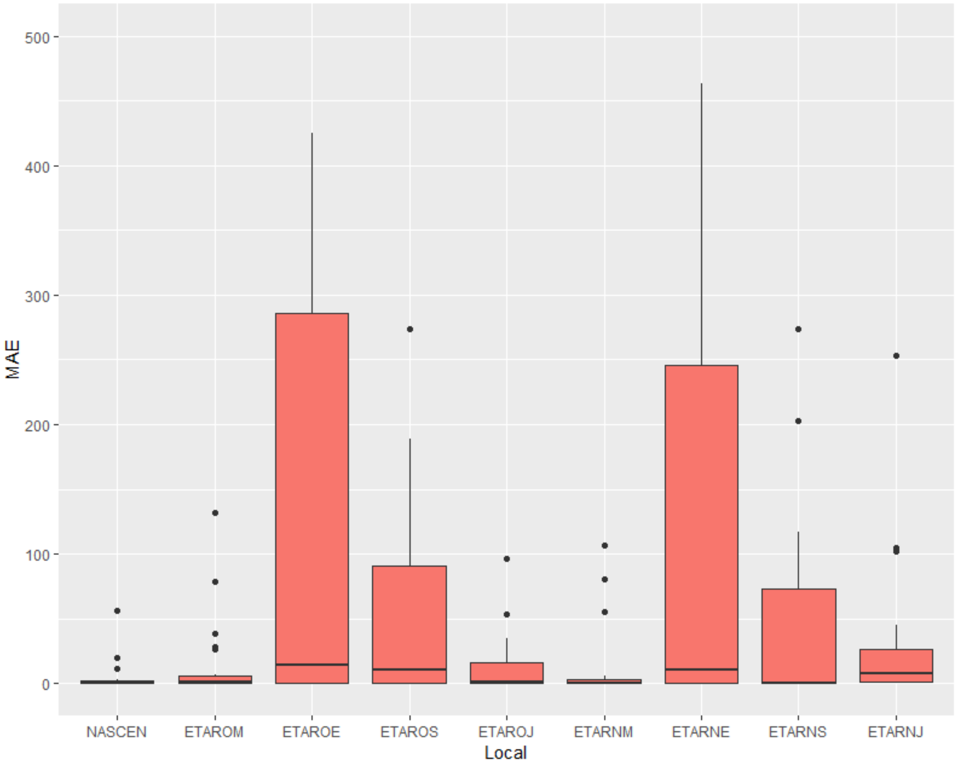


Figure A.6: Box-plots of the MAE estimates obtained using ETS and Top-down approach.

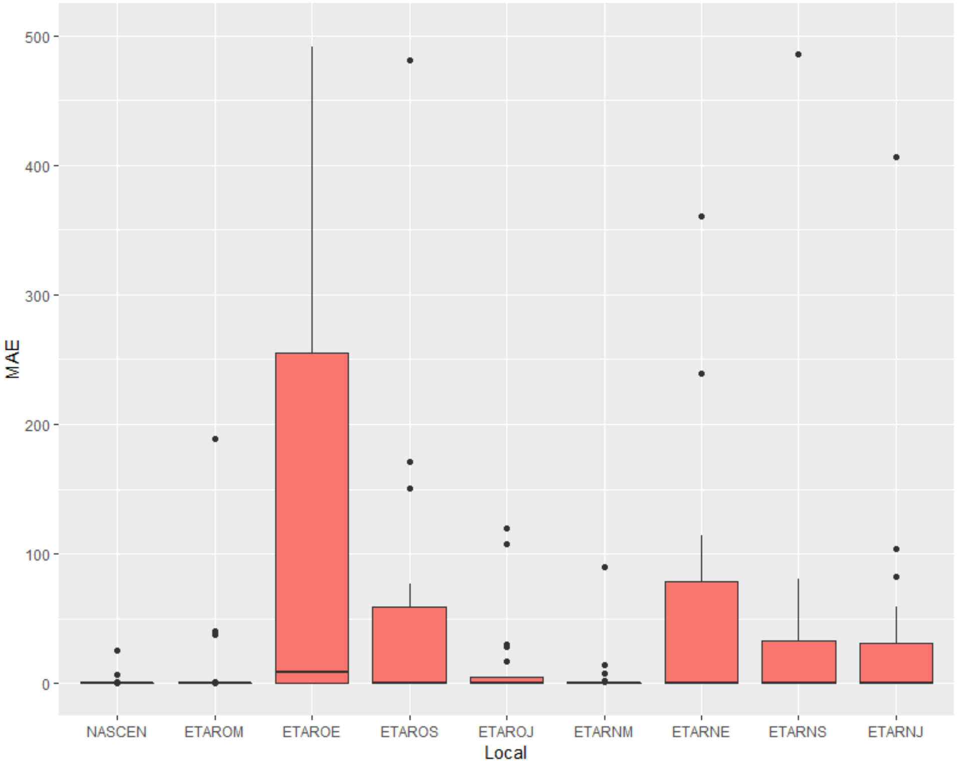


Figure A.7: Box-plots of the MAE estimates obtained using Random Walk and Middle-out approach.

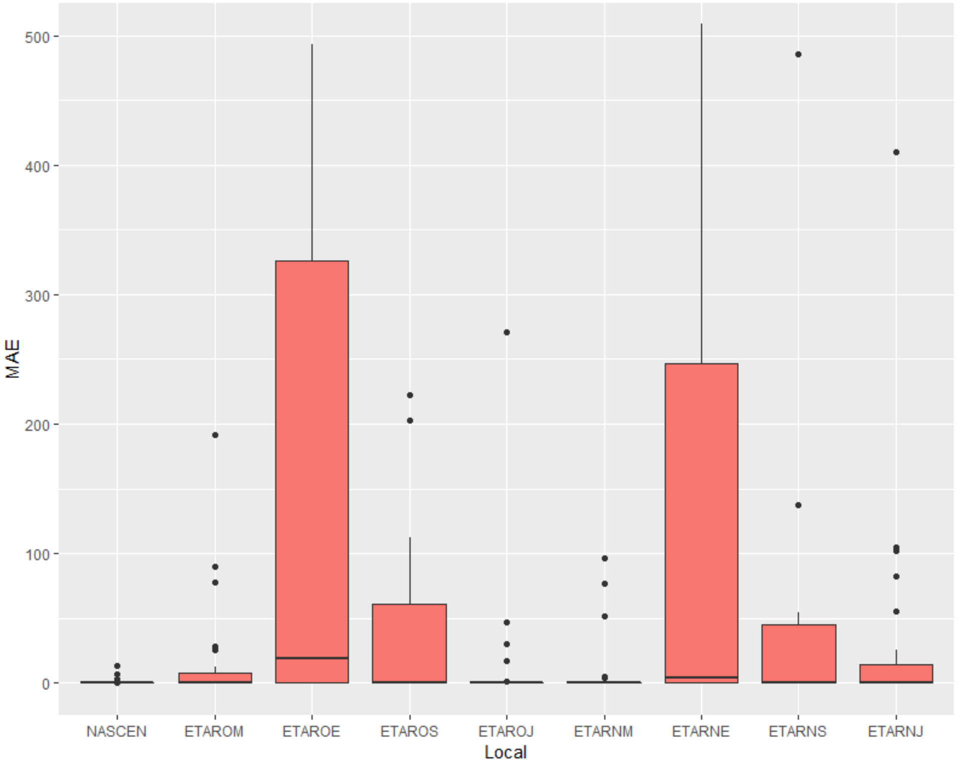


Figure A.8: Box-plots of the MAE estimates obtained using Arima and Middle-out approach.

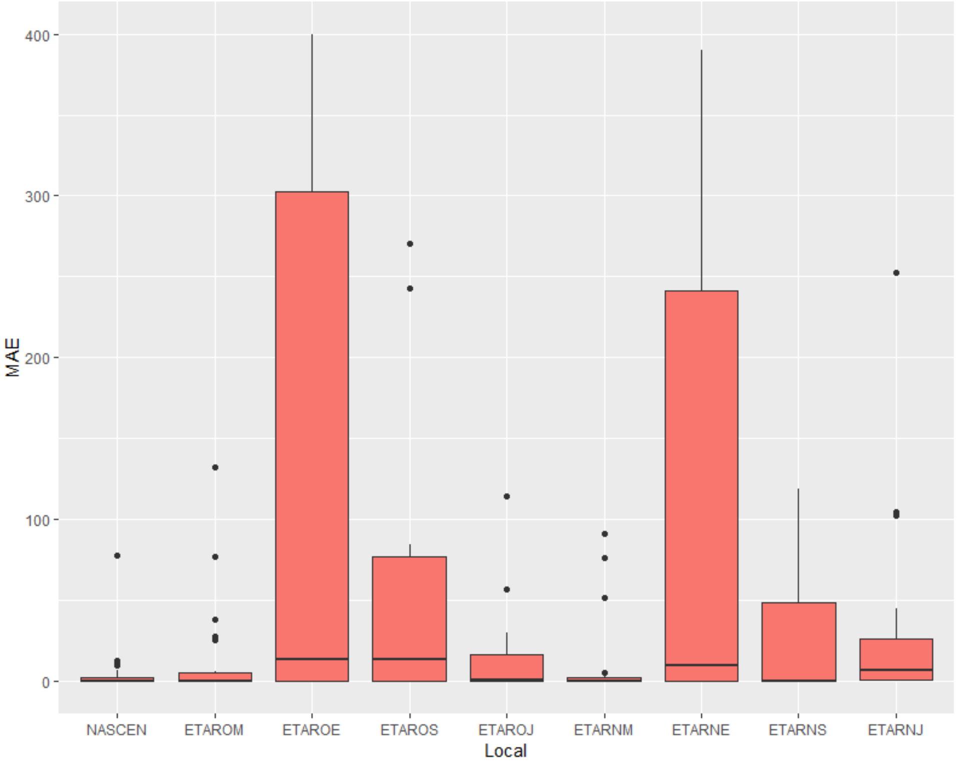


Figure A.9: Box-plots of the MAE estimates obtained using ETS and Middle-out approach.

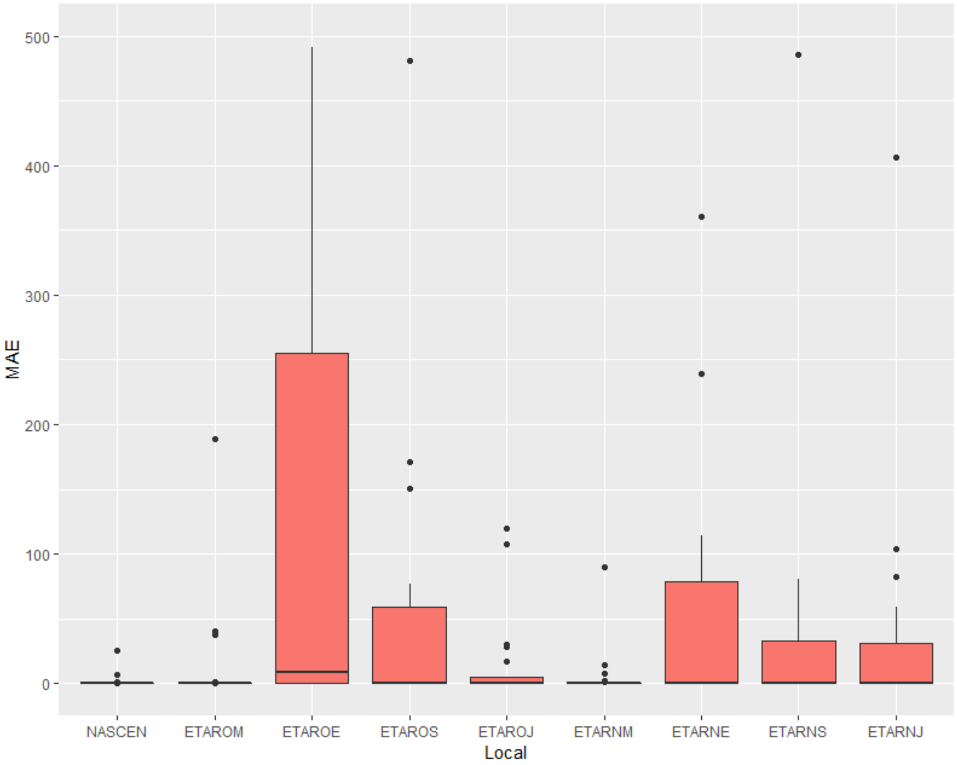


Figure A.10: Box-plots of the MAE estimates obtained using Random Walk and Bottom-up approach.

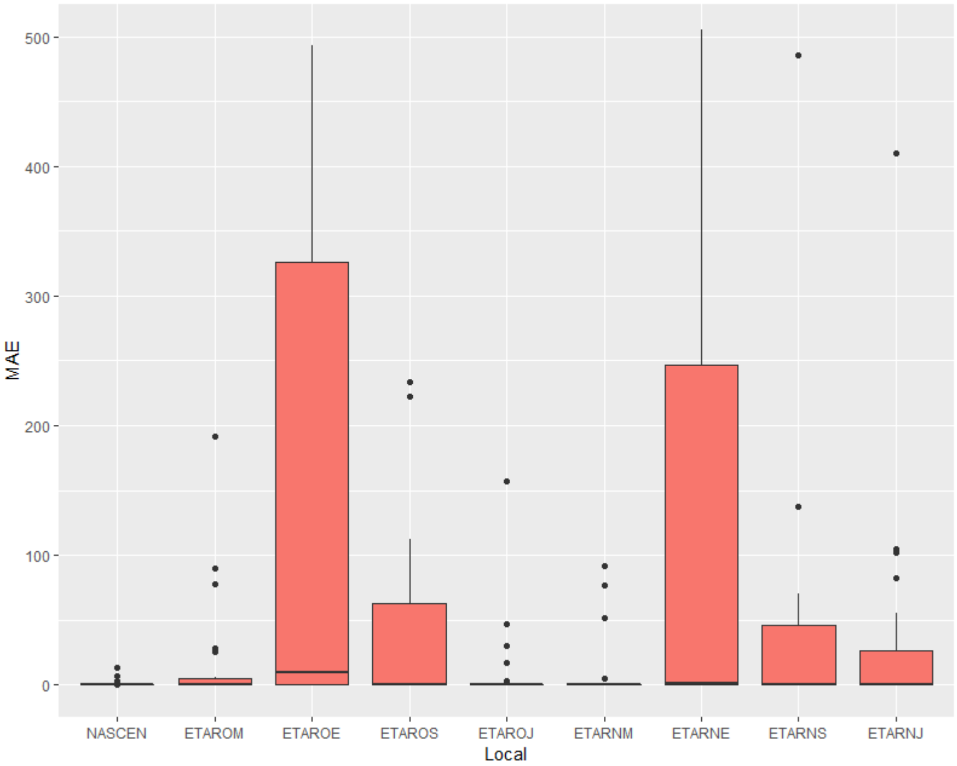


Figure A.11: Box-plots of the MAE estimates obtained using Arima and Bottom-up approach.

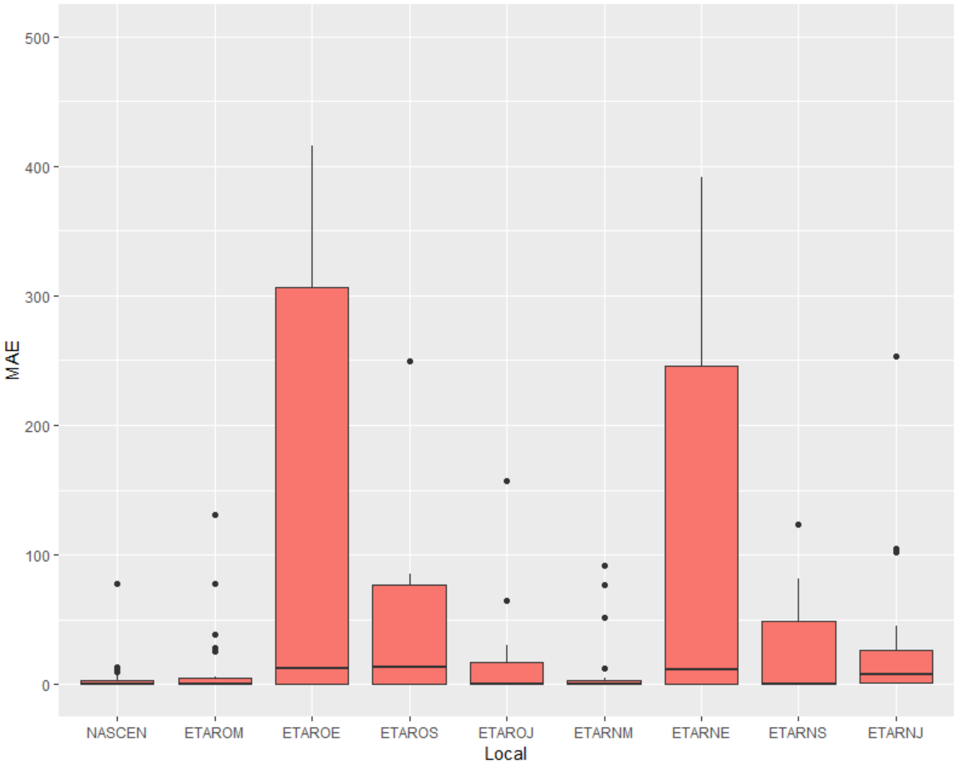


Figure A.12: Box-plots of the MAE estimates obtained using ETS and Bottom-up approach.

# Appendix B

# Case Study 2

In this appendix, we can see the remaining plots for the time series of the Caffeine and Lithium data set.

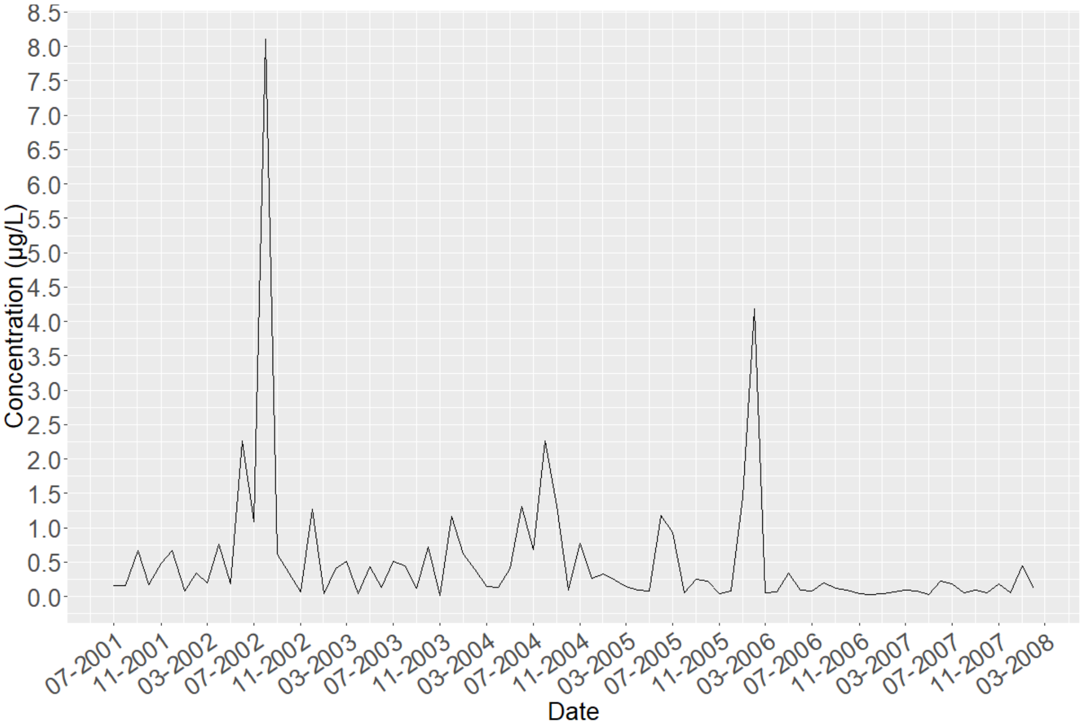


Figure B.1: Concentrations of the Caffeine 2 time series.

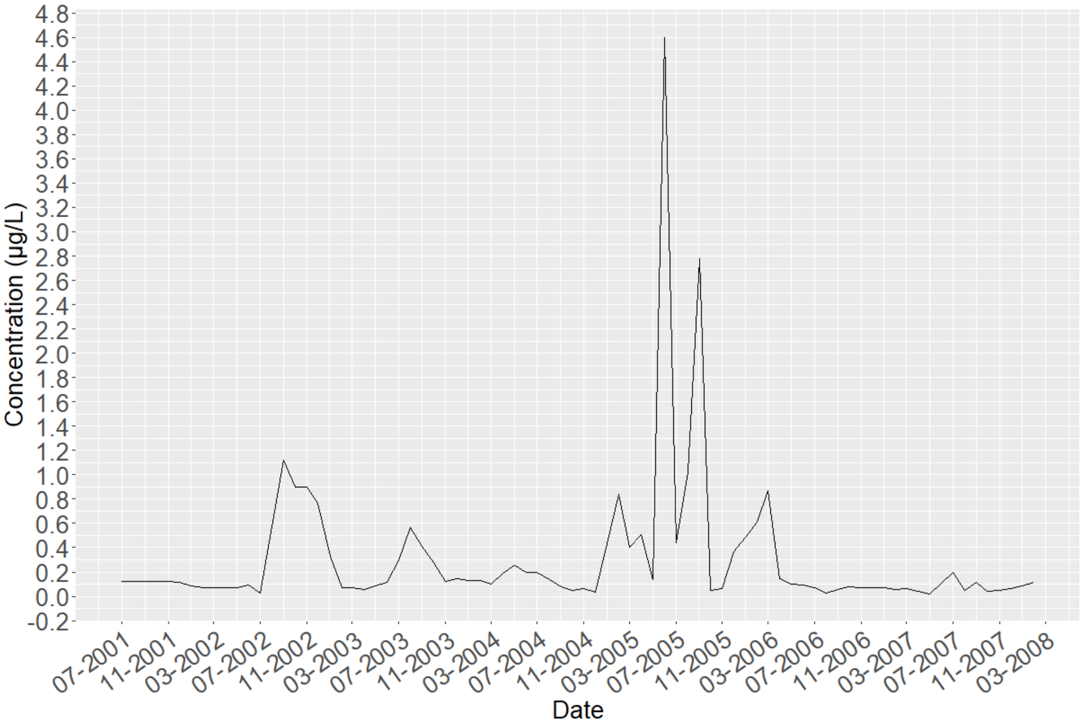


Figure B.2: Concentrations of the Caffeine 3 time series.

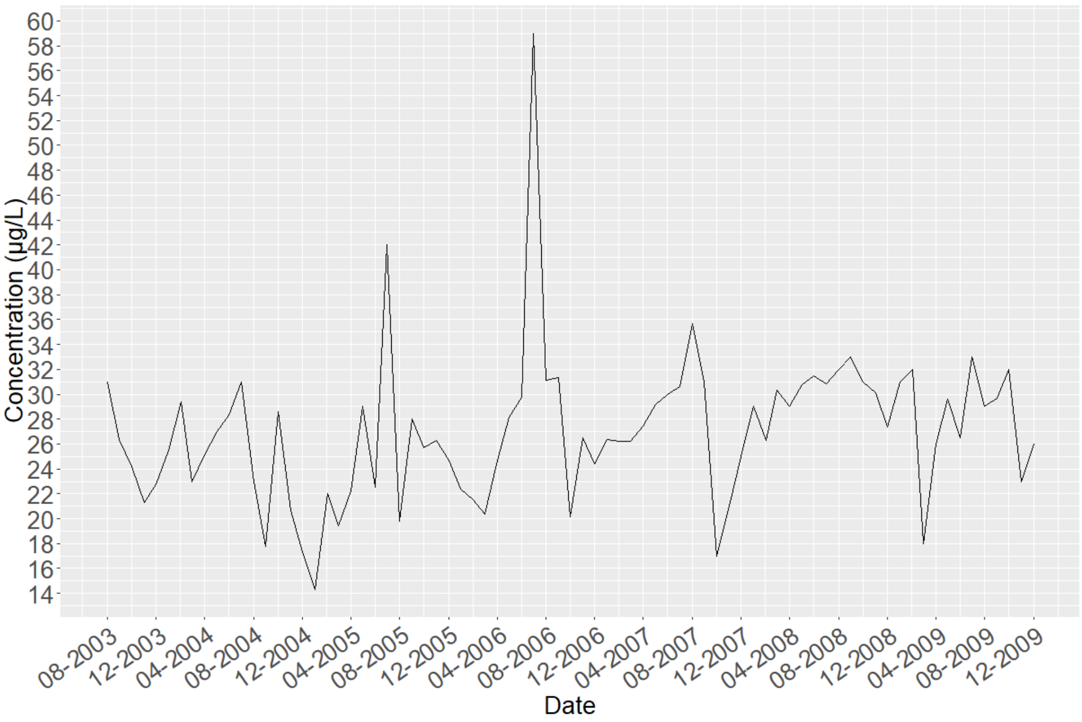


Figure B.3: Concentrations of the Lithium 2 time series.

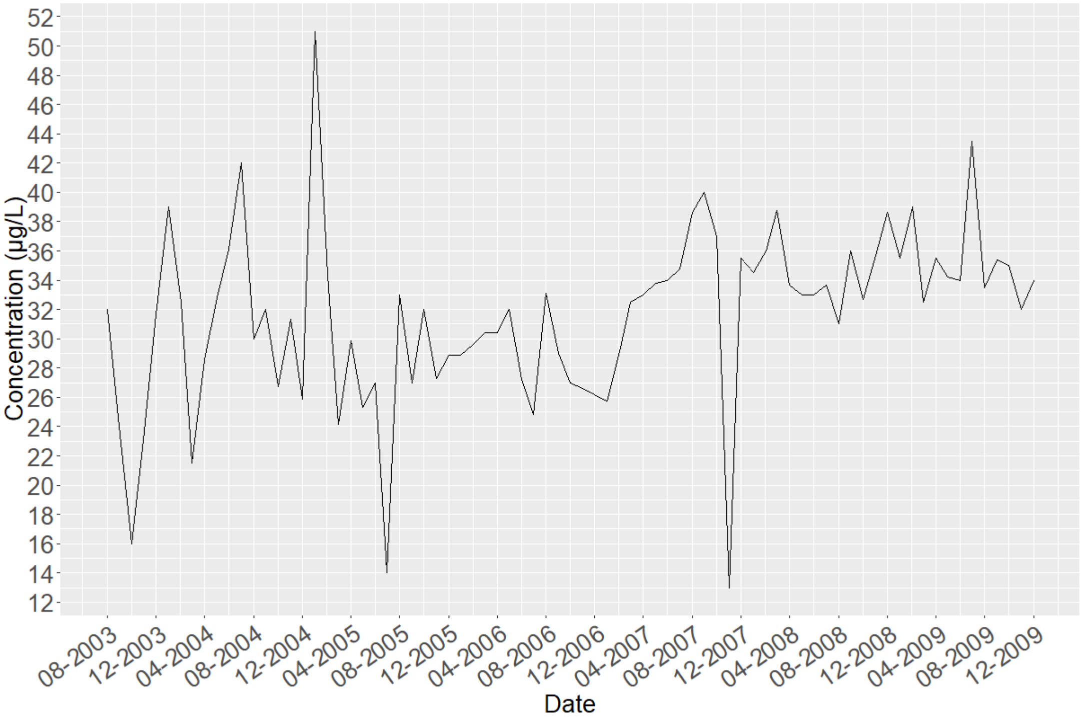


Figure B.4: Concentrations of the Lithium 3 time series.

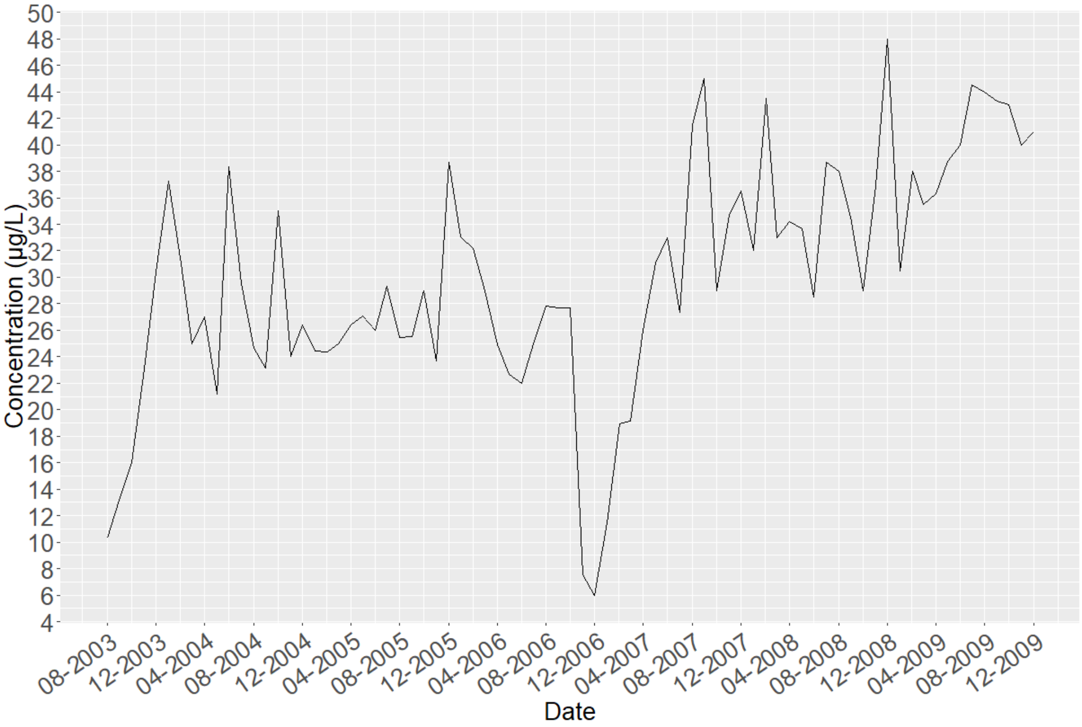


Figure B.5: Concentrations of the Lithium 4 time series.

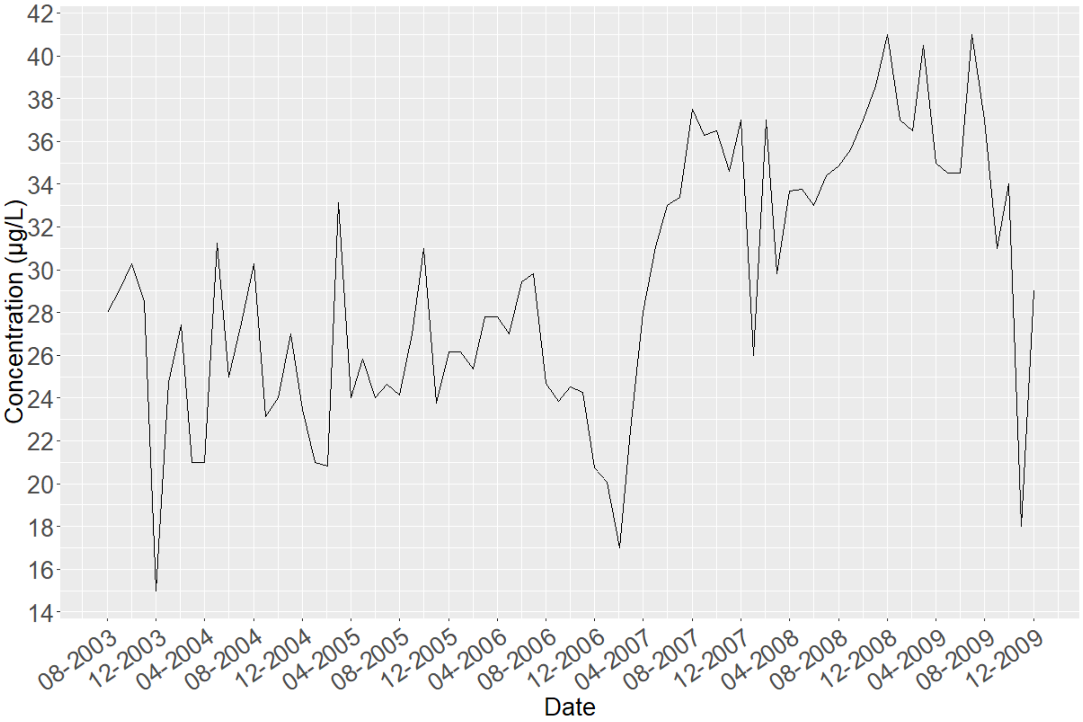


Figure B.6: Concentrations of the Lithium 5 time series.

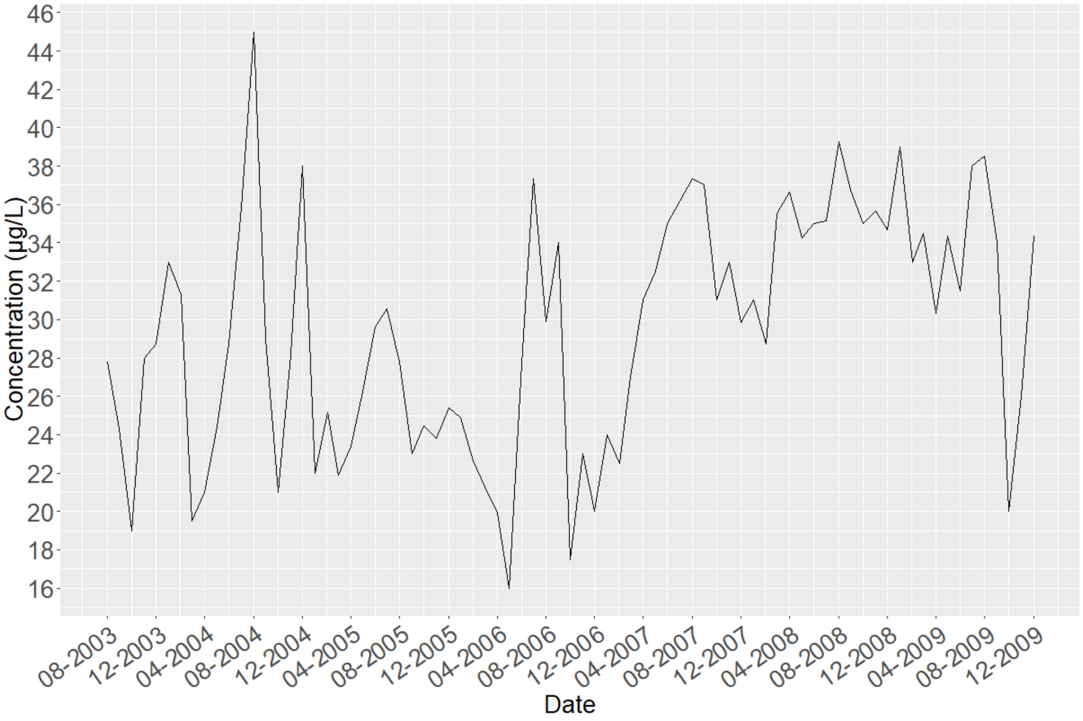


Figure B.7: Concentrations of the Lithium 6 time series.

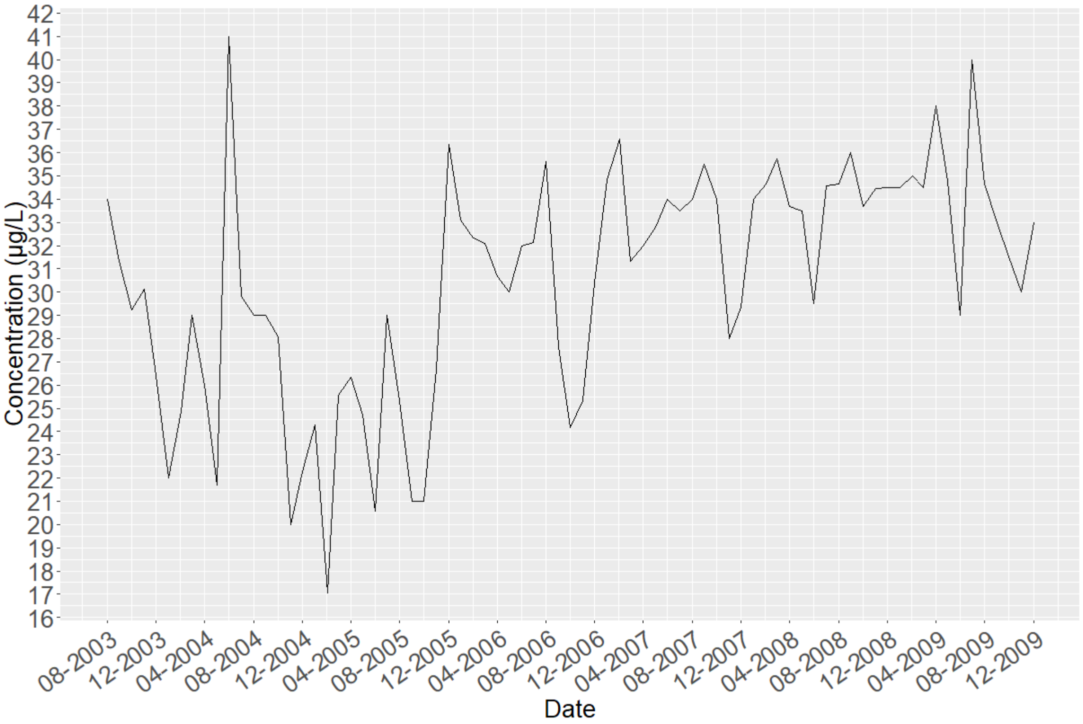


Figure B.8: Concentrations of the Lithium 7 time series.

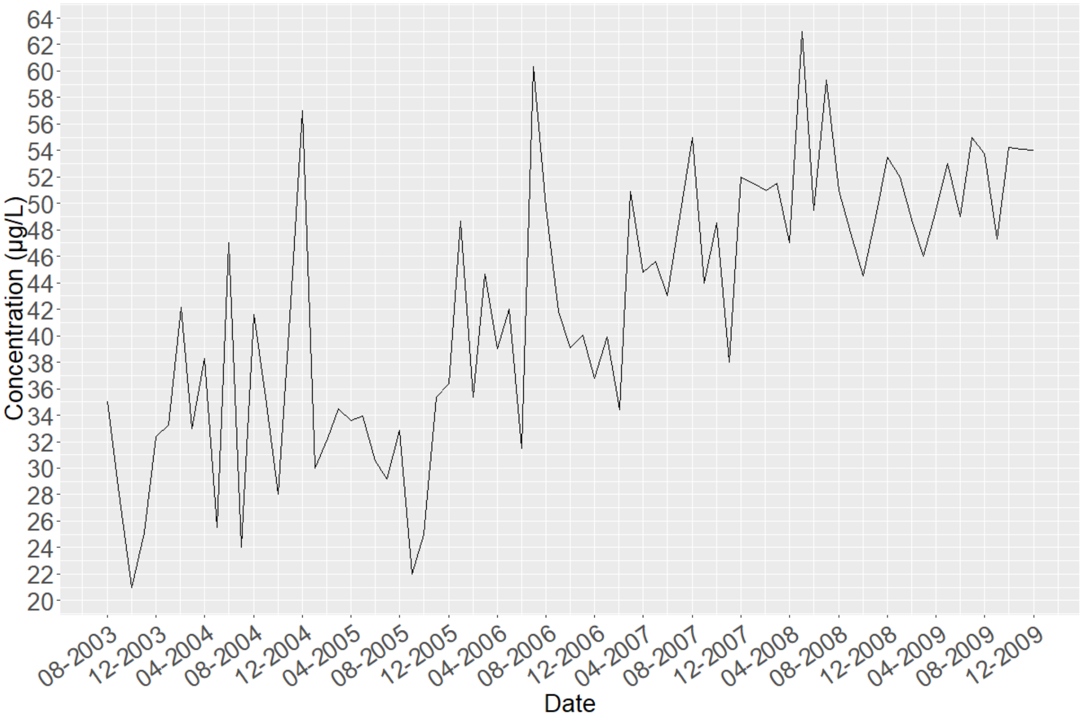


Figure B.9: Concentrations of the Lithium 8 time series.

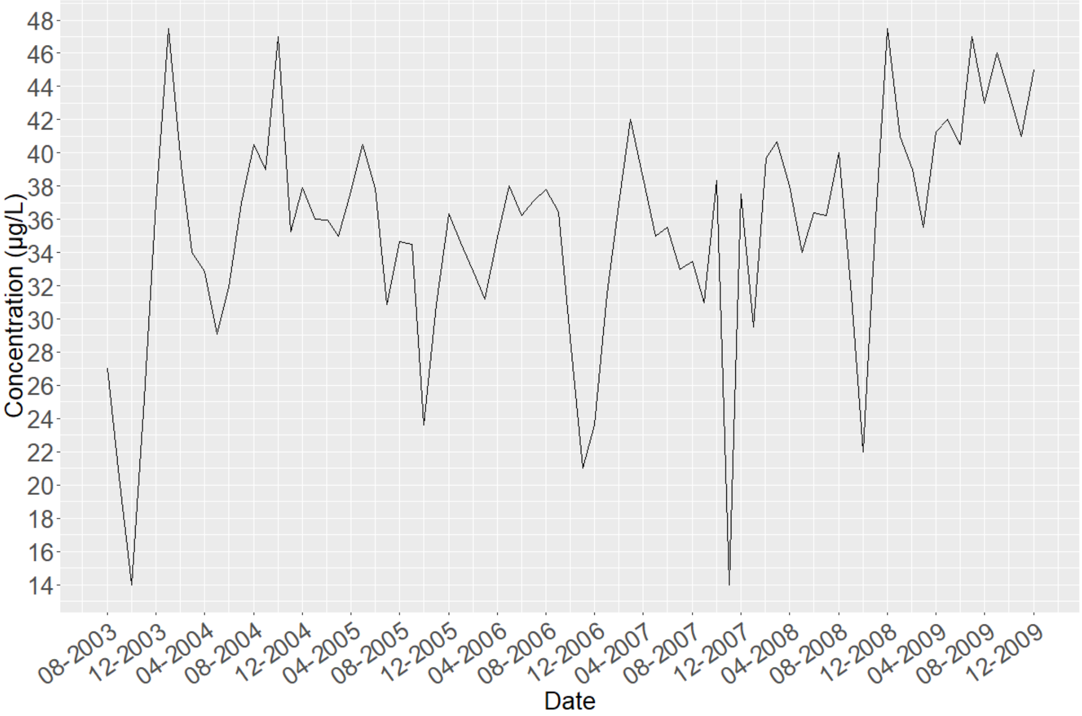


Figure B.10: Concentrations of the Lithium 9 time series.

We can also see remaining box-plots of the MAE estimates for the second case study.

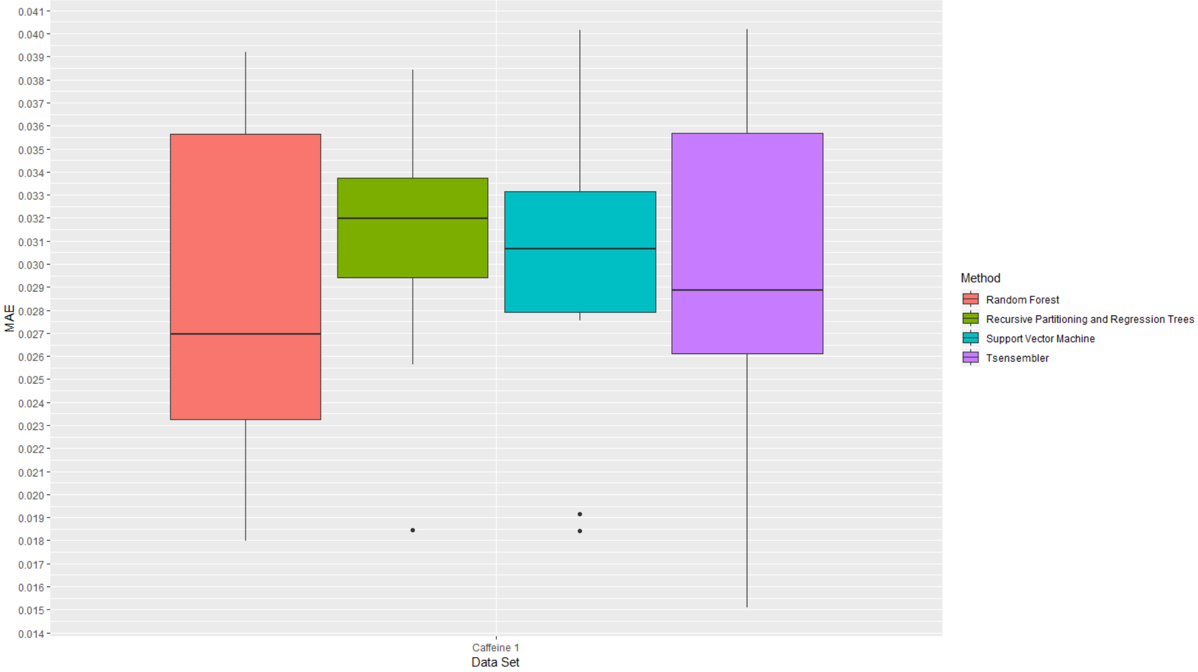


Figure B.11: Box-plots for the MAE estimates obtained in the prediction models for the caffeine 1 embedded time series.

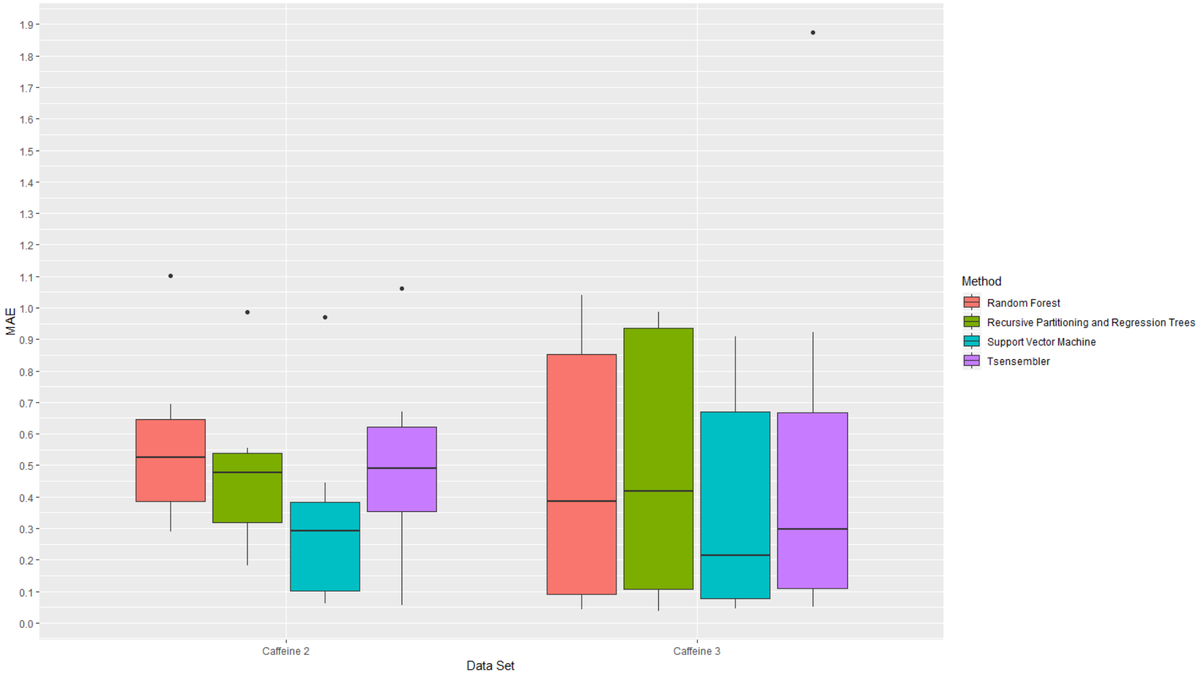


Figure B.12: Box-plots for the MAE estimates obtained in the prediction models for the caffeine 2 and caffeine 3 embedded time series.

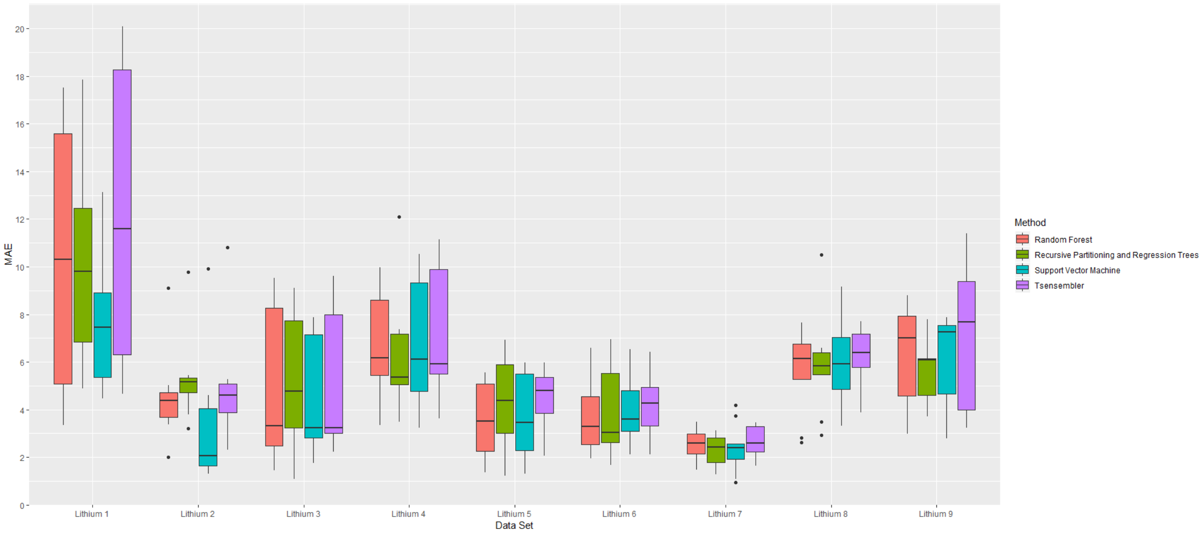


Figure B.13: Box-plots for the MAE estimates obtained in the prediction models for the embedded Lithium data set

# Bibliography

R. Adhikari and R. Agrawal. An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*, 2013.

I. Alon, M. Qi, and R. J. Sadowski. Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8(3): 147–156, 2001.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

V. Blüml, M. D. Regier, G. Hlavin, I. R. Rockett, F. König, B. Vyssoki, T. Bschor, and N. D. Kapusta. Lithium in the public water supply and suicide mortality in texas. *Journal of Psychiatric Research*, 47(3):407 – 411, 2013. ISSN 0022-3956. doi: https://doi.org/10.1016/j.jpsychires.2012.12.002. URL http://www.sciencedirect.com/science/article/pii/S002239561200372X.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

G. E. Box and G. M. Jenkins. *Time series analysis: forecasting and control, revised ed.* Holden-Day, 1976.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

B. W. Brooks, T. M. Riley, and R. D. Taylor. Water quality of effluent-dominated ecosystems: ecotoxicological, hydrological, and management considerations. *Hydrobiologia*, 556(1):365–379, 2006.

L. M. Burke. Caffeine and sports performance. *Applied Physiology, Nutrition, and Metabolism*, 33(6):1319–1334, 2008.

V. Cerqueira, L. Torgo, F. Pinto, and C. Soares. Arbitrated ensemble for time series forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 478–494. Springer, 2017.

G. Christakos. *Modern spatiotemporal geostatistics*, volume 6. Oxford University Press, 2000.

R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, 6(1):3–73, 1990.

J. M. Conley, S. J. Symes, M. S. Schorr, and S. M. Richards. Spatial and temporal analysis of pharmaceutical concentrations in the upper tennessee river basin. *Chemosphere*, 73(8): 1178–1187, 2008.

A. M. M. da Silva. Modelos preditivos aplicados ao retalho. Master's thesis, Faculty of Economics of University of Porto, 2015.

A. M. De Livera, R. J. Hyndman, and R. D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

F. Degenhardt, S. Seifert, and S. Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 2017.

T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

A. A. Dragulescu, M. A. A. Dragulescu, and R. Provide. Package 'xlsx'. *Cell*, 9:1, 2018.

F. Edition. Guidelines for drinking-water quality. *WHO chronicle*, 38(4):104–8, 2011.

R. K. Elissavet. *Missing Data in Time Series and Imputation Methods*. University of the Aegean, Department of Mathematics, February 2017.

M. Evans and C. Frick. The effects of road salts on aquatic ecosystems. 2001.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

X. Flores Alsina et al. *Conceptual design of wastewater treatment plants using multiple objectives*. Universitat de Girona, 2008.

FOTOCATGRAF. Fotocatgraf – graphene-based semiconductor photocatalysis for a safe and sustainable water supply: an advanced technology for emerging pollutants removal. http://www.fc.up.pt/fotocatgraf/. (Accessed on 07/02/2018).

E. S. Gardner Jr and E. McKenzie. Forecasting trends in time series. *Management Science*, 31 (10):1237–1246, 1985.

M. Gavrilescu, K. Demnerová, J. Aamand, S. Agathos, and F. Fava. Emerging pollutants in the environment: present and future challenges in biomonitoring, ecological risks and bioremediation. *New biotechnology*, 32(1):147–156, 2015.

V. Geissen, H. Mol, E. Klumpp, G. Umlauf, M. Nadal, M. van der Ploeg, S. E. van de Zee, and C. J. Ritsema. Emerging pollutants in the environment: a challenge for water resource management. *International soil and water conservation research*, 3(1):57–65, 2015.

L. O. Hall, K. W. Bowyer, R. E. Banfield, D. Bhadoria, W. P. Kegelmeyer, and S. Eschrich. Comparing pure parallel ensemble creation techniques against bagging. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 533–536. IEEE, 2003.

J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

D. J. Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

Z. He, X. Wen, H. Liu, and J. Du. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, 509:379–386, 2014.

Heidelberg University - NCWQR. Tributary data download – national center for water quality research. https://ncwqr.org/monitoring/data/. (Accessed on 03/06/2018).

M. Helbich, M. Leitner, and N. D. Kapusta. Lithium in drinking water and suicide mortality: interplay with lithium prescriptions. *The British Journal of Psychiatry*, 207(1):64–71, 2015.

R. M. Hirsch and L. A. De Cicco. User guide to exploration and graphics for river trends (egret) and dataretrieval: R packages for hydrologic data. Technical report, US Geological Survey, 2015.

M. Hofmann. Support vector machines—kernels and the kernel trick. *Notes*, 26, 2006.

C. C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5 – 10, 2004. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2003.09.015. URL http://www.sciencedirect.com/science/article/pii/S0169207003001134.

R. J. Hyndman. Forecasting: Principles & Practice. https://robjhyndman.com/uwafiles/fpp-notes.pdf. (Accessed on 03/04/2018).

R. J. Hyndman, Y. Khandakar, et al. *Automatic time series for forecasting: the forecast package for R*. Number 6/07. Monash University, Department of Econometrics and Business Statistics, 2007.

R. J. Hyndman, G. Athanasopoulos, and H. L. Shang. hts: An r package for forecasting hierarchical or grouped time series. *R Package (Available at http://cran. unej. ac. id/web/packages/hts/vignettes/hts. pdf)*, 2015.

M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15 (1):276, 2014.

K.-j. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55 (1-2):307–319, 2003.

D. T. Larose and C. D. Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Libesa. Using decomposition to improve time series prediction. https://quantdare.com/decomposition-to-improve-time-series-prediction/, September 2014. (Accessed on 09/03/2018).

C. F. Marlene Evans. The effects of road salts on aquatic ecosystems. https://brage.bibsys.no/xmlui/bitstream/id/201102/the_effects_road_salts.pdf, August 2001.

D. Meyer and F. T. Wien. Support vector machines. *R News*, 1(3):23–26, 2001.

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, and M. D. Meyer. Package 'e1071', 2018.

S. Moritz and T. Bartz-Beielstein. imputets: time series missing value imputation in r. *The R Journal*, 9(1):207–218, 2017.

NIST. Nist/sematech e-handbook of statistical methods. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm. (Accessed on 18/04/2018).

OpenCV. Introduction to support vector machines — opencv 3.0.0-dev documentation. https://docs.opencv.org/3.0-beta/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html. (Accessed on 09/02/2018).

P. Paíga, L. H. Santos, S. Ramos, S. Jorge, J. G. Silva, and C. Delerue-Matos. Presence of pharmaceuticals in the lis river (portugal): sources, fate and seasonal variation. *Science of the Total Environment*, 573:164–177, 2016.

J. Portilla. A beginner's guide to neural networks with r! https://www.kdnuggets.com/2016/08/begineers-guide-neural-networks-r.html. (Accessed on 09/20/2018).

J. S. Racine. Rstudio: A platform-independent ide for r and sweave. *Journal of Applied Econometrics*, 27(1):167–172, 2012.

G. A. Rob J Hyndman. Forecasting: Principles and practice. https://otexts.org/fpp2/hts.html. (Accessed on 09/20/2018).

J. Roberts, A. Kumar, J. Du, C. Hepplewhite, D. J. Ellis, A. G. Christy, and S. G. Beavis. Pharmaceuticals and personal care products (ppcps) in australia's largest inland sewage treatment plant, and its contribution to a major australian river during high and low flow. *Science of the total environment*, 541:1625–1637, 2016.

S. Sauvé, K. Aboulfadl, S. Dorner, P. Payment, G. Deschamps, and M. Prévost. Fecal coliforms, caffeine and carbamazepine in stormwater collection systems in a large urban area. *Chemosphere*, 86(2):118–123, 2012.

F. R. Spellman. *Handbook of water and wastewater treatment plant operations*. CRC press, 2013.

P. L. Spence. Using caffeine as a water quality indicator in the ambient monitoring program for third fork creek watershed, durham, north carolina. *Environmental health insights*, 9: EHI–S19588, 2015.

R. C. Team. Package “parallel.”. *R Foundation for Statistical Computing. Retrieved*, 18, 2013.

R. C. Team et al. R: A language and environment for statistical computing. 2013.

T. M. Therneau, B. Atkinson, and M. B. Ripley. The rpart package, 2010.

N. S. Thomaidis, A. G. Asimakopoulos, and A. Bletsou. Emerging contaminants: a tutorial mini-review. *Global NEST Journal*, 14(1):72–79, 2012.

L. Torgo. *Data mining with R: learning with case studies*. Chapman and Hall/CRC, 2016.

E. Turban, R. Sharda, J. E. Aronson, and D. King. *Business intelligence: A managerial approach*. Pearson Prentice Hall Upper Saddle River, NJ, 2008.

USGS Portal. Water quality portal - usgs. https://www.waterqualitydata.us/. (Accessed on 29/11/2017).

J. Warren, M. Fuentes, A. Herring, and P. Langlois. Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics*, 68(4):1157–1167, 2012.

M. West and J. Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

H. Wheater, R. Chandler, C. Onof, V. Isham, E. Bellone, C. Yang, D. Lekkas, G. Lourmas, and M.-L. Segond. Spatial-temporal rainfall modelling for flood risk estimation. *Stochastic Environmental Research and Risk Assessment*, 19(6):403–416, 2005.

H. Wickham. ggplot2: elegant graphics for data analysis. *J Stat Softw*, 35(1):65–88, 2010.

P. R. Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

G. P. Zhang and M. Qi. Neural network forecasting for seasonal and trend time series. *European journal of operational research*, 160(2):501–514, 2005.