TÉCNICO LISBOA

Water Quality Analysis in Water Distributions Systems using Multivariate Statistics and Water Quality Indexes

Rodrigo Mendes Ribeiro

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisors: Prof. Maria da Conceição Esperança Amado
Dr. Laura Sofia Pereira Pinto Monteiro

Examination Committee

Chairperson: Prof. António Manuel Pacheco Pires
Supervisor: Prof. Maria da Conceição Esperança Amado
Members of the Committee: Prof. Nelson Jorge Gaudêncio Carriço
Prof. Rui Henriques

November 2021

Acknowledgments

I would like to thank firstly to both my supervisors Prof. Conceição Amado and Dr. Laura Monteiro for the guidance along the past months of work as well as all the knowledge shared, tips and suggestions.

I would also like to thank my parents and friends for all the unconditional support that allowed me to keep challenging and pushing myself to do better along these 5 years of study.

I would also like to thank the financial support of Fundação para a Ciência e Tecnologia (FCT Portugal), through the research project DSAIPA/DS/0089/2018 (Water Inteligence System Data).

Abstract

Over the years water utilities have gathered information regarding the state of the quality of the water in its distribution systems. This quality is evaluated by a number of water parameters tested along the year in different sites, assuring the quality of the water provided to the customers is in conformity to the legislation. Investigating three case studies (Barreiro, Beja and Infraquinta), the goal is to perceive relations between the water parameters, detecting correlations and trends in the parameters variations' over time and along the water network. Another objective was to quantify this quality through water quality indexes (WQIs), where three newly created indexes are proposed. In order to detect simple correlations between the parameters the Pearson correlation matrix is used. As it has become popular in water distribution system analysis, an unsupervised Artifical Neural Network (ANN) class, the Kohonen self-organizing maps (SOMs) are a data mining tool being used in this field that allows a better understanding and a clearer view of the data through dimensionality reduction of the feature space to a 2D plane keeping the topology of the original data. Some other unsupervised methods such as Principal Component Analysis (PCA) and clustering techniques (KMeans and Hierarchical Clustering) were employed. As to the WQIs, two of them are adaptations from already existing work and the third one is a completely new approach using differences of the actual concentration of the parameter to its parametric value. The results obtained provided further insight of the water quality in each network, concluding that the water quality is of a very high standard in all cases.

Keywords

Data Mining, Water quality, Multivariate Statistics, Clustering, Self-organizing Maps (SOMs), Water Quality Indexes (WQI).

Resumo

Ao longo dos anos as entidades gestoras têm reunido informação acerca da qualidade da água nos seus sistemas de distribuição. Esta qualidade é avaliada por um número de parâmetros recolhidos ao longo do ano em diversos pontos da rede para assegurar que a qualidade da água fornecida aos consumidores está em conformidade com a legislação. Investigando três casos de estudo (Barreiro, Beja e Infraquinta), o objectivo é perceber as relações entre os parâmetros da qualidade da água, detetando correlações e tendências na variação dos parâmetros ao longo do tempo. Outro objetivo é quantificar a qualidade da água através de Indíces de Qualidade da Água (WQIs), onde três novos indíces são propostos. Para detetar a correlação entre os diferentes parâmetros é utilizada uma matriz de correlação de Pearson. Tem-se vindo a verificar um aumento no uso de uma classe de Redes Neurais Artificiais (ANNs) não supervisionadas. Os mapas de Kohonen ou *self-organizing maps* (SOMs) são uma ferramenta de data mining que permite uma melhor compreensão e uma visualização mais clara dos dados através de uma redução de dimensionalidade para um plano 2D mantendo a topologia dos dados originais. Outros métodos não superivisionados como Análise de Componentes Principais (PCA) e técnicas de *clustering* (K-Means e *clustering* hierárquico) foram aplicados. Dois dos WQIs utilizados são adaptações de índices de trabalhos existentes e o terceiro é um método de diferenças entre a concentração do parâmetro em questão e do seu respectivo valor paramétrico. Os resultados obtidos permitem um maior conhecimento acerca da qualidade da água em cada rede de distribuição, concluindo que a sua qualidade é de um nível muito alto em todos os casos estudados.

Palavras Chave

Data Mining, Qualidade da água, Estatística Multivariada, Análise de agrupamentos, Mapas auto-organizáveis (MAO), Índices de qualidade da água (IQA).

Contents

List of Figures

List of Tables

Acronyms

AHP Analytical Hierarchical Processes

ANNs Artificial Neural Networks

BMU Best-Matching Unit

EWQI Entropy weighted Water Quality Index

MDWQI Modified Drinking Water Quality Index

MIWQI Modified-Integrated Water Quality Index

PAHs Polynuclear Aromatic Hydrocarbons

PCA Principal Component Analysis

SOMs Self-Organizing Maps

TOC Total Organic Carbon

THMs Trihalomethanes

WQI Water Quality Index

1

Introduction

Contents

1.1 Motivation

Water is undoubtedly one of the most important resources for the survival of humans. Only about 2.5% of all the water resources on the planet is fresh and two thirds of this fresh water is located in the glaciers and ice caps. Removing out of the equation the water found in remote and inaccessible areas and the water derived from natural events (such as deluges and floods) which cannot be easily extracted, there is only 0.08% of all the fresh water on Earth [3] [4] that is used and exploited by humankind. Having access to fresh, clean, safe and drinking water is becoming scarcer by the day and it is limited for a part of the worldwide population. Thus, it is of our most interest to preserve and optimize this small percentage of fresh water available to us. With the growing uncertainties of global climate change and the long-term impacts of managements actions, the decision-making of how to make the best use of this asset is now more important than ever.

Water utilities in charge of treating and supplying drinking water are thus faced with the challenge of the sustainable and smart management of this precious resource. For that, water companies are investing in sensors, for better operational control, as well as in data analysis for the ever increasing amount of collected data. To answer this problem, among many other around the world, the WISDom project was created (Water Intelligence Systems Data project). This project aims to develop new algorithms and models that allow the extraction of relevant information from collected data. With the study of this data, the goal of the project is to support the decision-making of the entities and help them improve the operational management of their systems. This allows these organizations to conduct a more efficient, optimized and sustainable approach to their services.

This thesis, inserted in the WISDom project, aims to provide such solutions for three distinct entities partnered with the project. Infraquinta, located at Quinta do Lago, has a urban tipology and a floating population. Another partner is the Barreiro municipality (CM Barreiro), also with a urban tipology and about 80000 inhabitants. The last one is the municipal company of water and sanitation of Beja (EMAS Beja), with a rural tipology and a population on 34000 inhabitants [5].

In addition to the many continuously monitored parameters, such as pressure, flow rate and water levels in storage tanks, water utilities also collect many data regarding water quality. Water sampling and analysis is carried out routinely by all water utilities as imposed by national law. In order to comply with the legislation, the water utilities develop annual water quality control plans (in Portuguese, Plano de controlo de qualidade da água – PCQA) in which the sampling locations are identified along with the sampling frequency and the water quality parameters that must be analyzed. Every year, water utilities collect large numbers of water samples and store the correspondent results of the chemical and microbiological analysis, which are compiled and sent to the water services regulator (ERSAR - Entidade Reguladora dos Serviços de Águas e Resíduos). The data is analyzed regarding the conformity of the regulated water quality parameters and the number of analyzes effectively carried out, but no further

data analysis is made to the enormous data sets.

1.2 Objective

The objective of this thesis is to extract additional information from the water quality data collected by the water utilities over the years using artificial and machine learning methods that are are able to perceive relations between the parameters that are not easily identified.

The examination of the data is very important to detect possible correlations between parameters and to detect trends in the parameters variations over time and along the water network. Identifying correlations between parameters is important because it could happen that one parameter or a combination of several may allow a good estimation of another parameter, associated to a costly, difficult or long analysis in laboratory. For instance, while temperature and pH are easily measured on-site, other parameters, such as Trihalomethanes (THMs) require difficult and expensive analysis in laboratory to get accurate results. Another example is the microbiological analysis of colonies at 22℃ and 37℃, which take a lot of time to get the results.

When analyzing the relation between parameters, it is possible to identify which sample locations show similar parameter values, allowing us to establish a relation between different sites in the network, despite the fact they are close to each other or not.

Another goal of this thesis is to evaluate the water quality through water quality indexes. For this measure, three new proposed indexes were put into place to get a better understanding on how good the water quality is and which extra conclusions could be drawn.

1.3 Thesis Outline

The second Chapter is dedicated to the state of the art in neural networks and in water quality analysis and water quality indexes. The third Chapter is devoted to the background section, where the techniques used are explained. The fourth Chapter is related to data pre processing, where water quality is discussed and which steps are taken to process the data. The methodology section comes next, where it is explained how the methods are applied to analyze the water quality. The sixth Chapter concerns the results obtained for the three different case studies followed by conclusions drawn for each method applied. The last Chapter refers to the conclusion, where a brief conclusion of this thesis is done and a section of suggestions for future work to do.

2

State of the Art

Contents

One of the most important factors regarding a water distribution system is the quality of the water being delivered to every location. Most of the entities responsible for a water distribution network rely on periodic evaluations of the water quality parameters to ensure compliance with parametric values.

These entities have to guarantee that the quality of the water flowing into people's houses is in the best conditions possible, since the water in their network has to be eligible for human consumption. Another concern comes from the economic point of view since the responsible of the network prefers to identify any emerging problems beforehand instead of having to solve that complication once it has already happened which could cause great monetary expenditure to fix.

2.1 Neural Networks in Water quality Analysis

In addition to periodic analysis of water quality parameters, the entities are complementing this examinations investing in other methods that allow real time measures of water quality indicators such as residual chlorine. In order to understand the relation between the different parameters, Artificial Neural Networks (ANNs) have been employed to show in a simple way how parameters relate with each other, allowing the management entities to identify which parameters have an influence on others and how is that relation, and, if possible, make a prediction of those values [6] [7] [8] . In the domain of the ANNs, Self-Organizing Maps (SOMs), introduced by Kohonen [9] in 1990, have been the chosen method to illustrate how the parameters relate with each other in the study of the quality of the water. Interpreting these results allows conclusions to be drawn about the water network. Based on these facts, it is possible to optimize the network and improve its quality and/or detect any existent issue.

Overall, ANNs, namely SOMs, have been used as a very useful diagnosis tool to identify key areas and which parameters influence the quality of the water. In 2009, Wu et al. [10], through SOMs and K-Means cluster analysis, analyzed a urban water distribution network. The objective of this work was to make a comprehensive water quality evaluation system and a cluster analysis in this water supply network. This process was then used to evaluate the system performance. In their conclusion they stated that the network could be divided in 8 optimal cluster, each one containing a different average value for the water quality indicators, indicating a gradual decrease in water quality. This confirmed that SOM is very useful for water quality classification. In 2013, Juntenen et al. [7] analyzed samples collected from the Itkonniemi water treatment plant, located in Kuopio, Finland. The data was gathered during 951 days and had laboratory measurements of 41 different variables/parameters. Their objective was to classify, model and study data in the different stages of the water treatment process with SOM and posterior cluster analysis. It was concluded that this method facilitated data analysis and could be used to define process states and diagnose the behavior of the process in a convenient and user-friendly manner. It also enables efficient diagnosis in connection with a process, providing a clear illustration of

its condition and offered an applicable way of defining the best path for achieving a more efficient process with improved quality of water.

Since the 1990s the application of ANNs has been increasing, since they provide reliable and clear results regarding the prediction, forecasting and data processing in many domains. So, the introduction of ANNs to analyze water distributions system shouldn't be a surprise. In 2015, Wu et al. [11], delivered a protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. The analysis and evaluation of the models reviewed in this work was conducted based on scientific papers published between 2000 and 2012 in the field of ANNs implementation in water networks.

In 2007, Kalteh et al. [12] reviewed this growing ANN usage in the analysis, estimation and prediction of various hydrological processes, such as river flow and rainfall-runoff, precipitation, in surface water quality and in other climate and environment related processes. In the studies presented there, it is shown that in many cases, SOMs can outperform other methods to solve various issues in water resources and hydrology.

One application of SOMs regarding water distribution networks is to analyze the deterioration of the quality of the water, in order to improve the water network and increase the water quality being provided to customers. Regarding this matter, Mounce et al. [13] analyzed the biofilms attached to to the inner pipe wall of a test rig. Using Principal Component Analysis (PCA) and SOMs, the relationship between the water physico-chemical and microbiological characteristics was inspected. Also in the domain of Bioinformatics, the authors managed to conduct experiences and replicate the processes happening inside the pipes of the network. It was concluded from this work that PCA and SOMs proved to be a reliable and clear approach to examine further relations between variables and to explore hypothesis for increased understanding of water quality degradation in networks.

In 2014, Mounce et al. [14] discussed the relation between stagnation of the water and its consequent deterioration and aimed to describe associations between deterioration of water quality through proxy measurements (high iron, manganese and turbidity) and stagnation in water distribution systems. The case studied referred to a water supply zone for a town in the UK with both rural and urban areas of about 5000 customers. Information regarding the different water parameters of interest, the pipes infrastructure, the hydraulics model and customer and water service contacts were made available. After applying the SOM method and interpreting the results, the authors concluded that the risk for water quality stagnation was found to be greatest in cast iron pipes with medium diameters, medium to high residence times, high water age, high condition code and located in rural areas.

The accumulation of particles in potable water distribution systems can lead to the discoloration of the water which can be harmful for the final customers. The stocking of this material, which can contain high levels of metal, organic/inorganic compounds and micro-organisms risky for people's health, occurs in

the pipe's surface and acts as cohesive layers. Once mobilized, these particles lead to the discoloration of the water and the consequent decrease in its quality. The accumulation rate of this material was the subject of the work by Mounce et al. [15]. Employing techniques such as the EPR (evolutionary polynomial regression) and SOMs in three different data sets, different conclusions were drawn. The first case is an UK case study of a network serving 40 million customers, with detailed information about the pipes in the network, water quality factors, water treatment processes and site-specific details. The second case scenario, is a dutch local scale long-term monitoring data set. It was collected from an area in Purmerend, Netherlands with 2310 connected habitations. Information about water supply and treatment processes and pipe material were available. The third and last case study is a dutch local scale highly repeated flushing case. Collected from an area in Volendam, Netherlands, the same information from the other dutch case was made available to perform this study. The conclusions for the first case were very detailed and the regeneration rated were calculated with precision. However, this conclusion was very different from the other two data sets, which highlights the difference in the quality and quantity between the collected data of each entity. Through the methodology used, they managed to identify the main components responsible for the discoloration of the water in the network and the impact of the frequency of flushing (causing the mobilization of the accumulated particles).

Another factor responsible for the deterioration of the water quality can be water age and temperature. These elements were identified and submitted to a further study accomplished by Blokker et al. [16]. This paper reports on how various microbial parameters correlated with modelled water ages and were influenced by water temperatures in three drinking water distribution systems. Data sets from Dutch and UK water companies contained information regarding water age and pipe material taken from hydraulic models of the associated networks. The Dutch data set represented ten years of regulatory water quality sample results (over 25000 entries), whereas the UK data set was much smaller, but a more accurate hydraulic model was used and intensive water quality sampling was targeted at areas of the network containing specific water age volumes. The authors resorted to SOMs to analyze the different relationships between variables. It was concluded from this study that water age and temperature may be treated as independent parameters regarding their influence on the deterioration of the water quality, there is an apparent influence of temperature on *Aeromonas* and HPC (heterotrophic plate counts) at 22℃. The correlation with water age is smaller, and there is little added value in considering maximum rather than average modelled water age. The parameter of water age thus seems to be of little value as an indicator for specific microbial water quality.

One environmental and ecological problem regarding the quality of the water in networks is the pollution produced by humankind. The contamination of water by human activities can severely worsen the quality of the water delivered for human consumption. To address this matter and find key insights to solve this emergent problem, SOMs revealed to be an excellent tool to visualize and interpret underlying

relations. In 2015, An Yan et al. [8] analyzed the surface water quality in Tolo Harbour and Channel Water Control Zone in Hong Kong. Making use of PCA and SOMs, the authors managed to identify key areas that contributed to the deterioration of the quality in that zone. In this data set there were 4752 measurements of twelve parameters collected between 2009 and 2011. It was concluded that there were results in monitoring stations being affected by untreated sewage effluents and the methods utilized showed relations between some of the parameters of interest. Also, they managed to classify the data in 4 different clusters that describe the normal condition of the study in the area.

Speight et al. [17] resorted to SOMs to identify the dominant mechanisms of iron release in a three water supply networks in the UK. This release and the combination of organic and inorganic compounds at sufficient concentrations, affect the turbidity of the water, resulting in the discoloration of the water, affecting its quality. Discovering why and where this occurs is therefore of great importance. The first water distribution system provided by company A serves 5.4 million people with 245 water treatment works. Data from this company was collected between January 2012 and May 2016, containing information about the water composition, its parameters, the treatment processes and the configuration of the network. The second case study were three cities within Company B. The three cities served a population of 142000, 199000 and 261000. The last case, company C serves 3.1 million people, with 63 water treatment works. It provided data from January 2008 to September 2014, not containing any information regarding the hydraulic model and pipe details. In this work, using SOMs, it was easily identified how the parameters, namely iron, related to the region where the samples were taken, the disinfection used there and pipe material in that area. It was concluded that SOM analysis offered a simple and straightforward way to capture relations between the parameters (such as iron, manganese and turbidity) and to demonstrate the strength of trends and multivariate correlations. It was also stated that cast iron pipes were often the focus of discoloration and were associated to high levels of iron, even if some were performing well despite their age/condition. This analysis allowed companies to address this and other issues that cause iron release and discoloration in a water distribution system and therefore improve the quality of the water being delivered to their customers. A similar study was done by G. Kyritsakas et al. [18]. The goal was to identify the correlations between the main characteristics of water in both systems and investigate the change of the water quality in a system that switches its disinfection type from chlorination to chloramination, using neural networks, such as SOMs, to highlight its potential in water quality overview. Using a UK Water Company data set that serves over 5 million people, containing information about water quality for a period between 2012 to 2018. The data set also included regulatory water quality results from water treatment works, the service reservoirs and the customers' taps. Data about pipes and additional info taken from the Water Company records, rainfall and Asset Management data was also at their disposal. The authors concluded that SOMs are a powerful tool to identify relations between the parameters and the disinfection method used. It was useful to iden-

tify sampling locations were the switch of disinfection methods had a more important impact and how that impact was measured. In other words, it allowed them to identify the locations which parameters suffered the most changes with the switch. This investigation also allowed to determine which areas required a more exhaustive investigation.

2.2 Water Quality Indexes as measure of Water Quality

Regarding the water quality index, several studies were of importance and influenced the way this thesis approached this question. In order to be able to evaluate the quality of the drinking water present in ones home, multiple entities and organizations of the sector searched for a way to rate the quality of the water based on its chemical, physical and microbiological parameters. Since the parameters always have a parametric value they cannot or should not pass, in order to comply with national legislation, one type of index, based on weights and on this limit value such as the Weight Arithmetic Index, made more sense to be used. A Water Quality Index (WQI) is an index that transform complex data regarding the values obtained after analyzing the water parameters in a certain network into an actual number or interval that categorizes the different samples and thus the entire network based on how good (or bad) the values of the parameters are, relatively to their limit values. Another great advantage of this kind of indexes is the fact one can utilize all available parameters to perform this quality assessment. The different indexes here described are quite similar among themselves and only differ in the calculation of the weights. The different authors that utilize these indexes may or may not deal with microbiological parameters, since their parametric value for human consumption is 0, and this causes a problem when calculating the weights associated to the index in question. The use of microbiological parameters to quantify the quality of the water is of the essence, even though, as it will be seen, not many authors include them in their works.

The information regarding each index can be summed in the following Table. Similar indexes are stated below each other. Water Arithmetic Indexes are compared with other indexes that evaluate the quality of the water. The microbiological (MB) column refers to the presence of one or more microbiological parameters in the calculation of the index. The notation used is: C_i is the concentration of the parameter i in that sample, K is a proportionality constant, S_i is the standard/parametric value of the parameter i, Q_i is the quality rating of the parameter i in that sample, W_i is the weight associated to the parameter i, n is the total number of samples where parameters were analyzed, C_0 is the ideal concentration of the parameter i in pure water -this only affects the pH and the dissolved oxygen parameters, where its values are 7.0 and 14.6, respectively - for the other parameters it takes the value 0, the WQI is the overall index of the network. According to how each index is calculated, the range of ratings that quantify the water quality is also different.

Notice, as well, that the difference on how Q_i is calculated depends on whether parameters such as pH and Dissolved Oxygen (DO) are involved when calculating the index. Either way, it was decided to show them as different indexes, since the formula is different if those parameters are present and thus stay truthful to the author's work. If these parameters are not present, the formula becomes the same, and utilizing one or the other is the same thing.

Table 2.1: Calculation of the different indexes

	Qi	K	Wi	WQI	MB	Reference
Index 1:	$\frac{C_i}{S_i} \times 100$	NA	$\frac{1}{S_i}$	$\sum^n \frac{Q_i W_i}{W_i}$	✗	J. Yisa et al. [19]
					✗	Rahman et al. [20]
Index 2	$\frac{C_i - C_0}{S_i - C_0} \times 100$	$\frac{1}{\sum^n (1/S_i)}$	$\frac{K}{S_i}$	$\sum^n \frac{Q_i W_i}{W_i}$	✗	Tyagi et al. [21]
					✗	Al-Alafify et al. [22]
		cst			✗	Bouslah et al. [23]
		$\frac{1}{\sum^n (1/S_i)}$	$\frac{1}{\sum^n S_i}$		✗	Dutta et al. [24]
Index 3	$\frac{C_i - C_0}{S_i - C_0} \times 100$	NA	$\frac{w_i^*}{\sum^n w_i^*}$	$\sum^n W_i Q_i$	✓	Ibrahim et al. [25]
Index 4	$\frac{C_i}{S_i} \times 100$	NA	$\frac{w_i^*}{\sum^n w_i^*}$	$\sum^n W_i Q_i$	✗	Badeenezhad et al. [26]
					✗	Olasoji et al. [27]
					✓	Alomran et al. [28]
					✓	Alver et al. [29]
Index 5 (RWQI)	Rating Curves	NA	$\frac{(1/w_i^*)}{\sum^n (1/w_i^*)}$	$\prod^n Q_i^{W_i}$	✓	Almeida et al. [30]
Index 6 (UWQI)	Rating Curves	NA	$\frac{w_i^*}{\sum^n w_i^*}$	$\sum^n W_i Q_i$	✓	Boyacioglu et al. [31]
Index 7 (OWQI)	Rating Curves	NA	NS	$\sum^n W_i Q_i$	✓	Cude et al. [32]
Index 8 (NSFWQI)	Rating Curves	NA	NS	$\prod^n Q_i^{W_i}$	✓	Cude et al. [32]
Index 9	Rating Curves	NA	NS	$\sqrt{\frac{n}{\sum^n (1/S_i^2)}}$	✓	Cude et al. [32]

NA values mean that for that index, K is not needed to calculate the index. NS means Not Specified in the paper where the index is discussed. For this case, in particular W_i are not specified for the last 3 indexes, and are only refered to as the weight given to subindex (or the corresponding Q_i value of the i parameter. The weight w_i^* is a relative, arbitrary weight assigned to each parameter. It can be determined by experts, by previous work in the area or even after a PCA analysis. It usually is a number in a scale from 1 to 5. The rating curves (they can appear under the subindex category also) are functions defined by experts and investigators that quantify the quality of a parameter translating its concentration value in a sample to a scale of 0 to 100, where 100 represents a very good water and 0 a very poor one. To perform this kind of operation, non-linear regression can be used.

Even though, some papers refer the importance of microbiological parameters, in some of them

these kind of parameters are not used when calculating the WQI. If this is the case, in Table 2.1, the microbiological column has a ✗. Keep in mind that for almost all documents or regulations that control the parametric values regulate that the microbiological concentration in drinking water can't be over 0, there are some that allow a slighter higher value (and thus not 0). This does not cause the problem discussed above where the standard/parametric value S_i is 0, causing an impossibility to calculate the rating Q_i. However, once more to stay truthful to the work developed by the authors, if in any way the microbiological parameters are used to explain the WQI, then they have a ✓in the respective column in Table 2.1.

Apart form the indexes described in Table 2.1, there are some others that require a more detailed explanation, due to its complexity.

Index 10 is referred to as the original CCME WQI. This index was established under the Canadian Council of Ministers of the Environment, hence the name. This index is described by Khan et al. [33]. It is further explained by Hurley et al. [34], where a modified version of this index is also developed. This index is based on calculating three different measures F_1, F_2 and F_3 and applying them in another formula given by:

$$\text{CCME WQI} = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \tag{2.1}$$

Mohebbi et al. [35] also worked in a similar index and making adjustments to the weights when calculating the different components, calling it Modified Drinking Water Quality Index (MDWQI). This index is also used by Oliveira et al. [36] to evaluate the water quality in a Brasilian Water Reservoir.

Two other types of indexes worth mentioning to evaluate the WQI are the Modified-Integrated Water Quality Index (MIWQI) and the Entropy weighted Water Quality Index (EWQI).

The first one is further explained by Islam et al. [37] and Amiri et al. [38]. In the second one, the authors mention different ways to build a matrix Y according to the final wished objective. Their intent in the work mentioned was to evaluate the water quality in a water network in Lenjanat, Iran. In both works, none microbiological parameters were utilized to build both indexes mentioned above.

Introducing the first one explained by Islam et al. [37], there are several steps:

1. Range $=$ TC $-$ RC, where TC is the Tolerable Concentration (Parametric Value for the parameter) and RC is the Required Concentration (in the case the parameter has a required minimum value)
2. MTC $=$ TC $-0.15\times$Range, where MTC stands for Modified Tolerable Concentration.
3. Calculation of the subindexes (SI) or quality rating:
    - If the value of the parameter i is above RC and below the MTC, the quality rating is $SI_1 = 0$
    - If the observed values of the parameter i are below RC, the quality rating is $SI_2 = \frac{RC - P_i}{RC}$
    - If P_i is above the MTC, the quality rating is $SI_3 = \frac{P_i - MTC}{MTC}$

where P_i is the water quality of the parameter i.

Finally, the MIWQI is computed as the sum of all all subindexes of each parameter acquired from step 3. The values of each parameter is thus calculated as:

$$MIWQI_i = \sum^{n} SI_{ij} \tag{2.2}$$

where SI_{ij} is the subindex of the sample i and the water parameter j. This index is applied in water quality assessment by Islam et al. [37] in a region of Bangladesh.

Regarding the EWQI, it can also be divided into different steps:

1. Create a matrix X:

   X = $\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mm} \end{bmatrix}$

   where $m(i = 1, 2, ..., m)$ denotes the total number of samples, and $n(j = 1, 2, ..., n)$ denotes the number of parameters in each sample and x_{ij} denotes the concentration of sample i for the parameter j.

2. Build a new matrix, Y, following a normalization construction function, and $y_{ij} = \frac{x_{ij} - x_{ij}^{min}}{x_{ij}^{max} - x_{ij}^{min}}$ where the values x_{ij} are the values from the matrix X, x_{ij}^{min} and x_{ij}^{max} are the lowest and highest values of the parameter in the samples.

3. Define $P_{ij} = \frac{y_{ij}}{\sum^{n} y_{ij}}$

4. Compute $e_j = \frac{1}{\ln m} \sum^{n} P_{ij} \ln P_{ij}$ where e_j is the information entropy

5. $\omega_j = \frac{1 - e_j}{\sum^{n} (1 - e_j)}$ where ω_j is the entropy weight

6. $q_j = \frac{C_j}{S_j} \times 100$ where q_j is the quality rating, C_j is the concentration of the parameter j, and S_j is the standard/parametric value for the parameter j.

   Finally, the index is calculated as:

$$EWQI = \sum^{n} \omega_j q_j \tag{2.3}$$

These are just some examples of indexes and there are a lot more to be explored. Some references to fuzzy inference systems based indexes can be found in Gharibi et al. [39]. Other approach using Analytical Hierarchical Processes (AHP) were demonstrated and furter explained by Sutadian et al. [40]. There was also an investigation on indexes regarding other domains apart from the water sector, two examples are the mine soil quality index [41] and diet quality index [42].

3

Background

Contents

3.1 Overview

In this thesis, three different data sets with different properties from different locations are analyzed: Barreiro, Beja and Infraquinta. The data sets contain information about the results of different water quality parameters. The Barreiro data set has the results collected from the network in the period between 2010 until 2019. The Beja data set has information collected from 2009 to 2019. And the Infraquinta data set has data from 2008 to 2019, with the exception of the year 2011.

For all three of them, the same type of analysis is made.

The years are all condensed together in a data set to have a general idea of the network in all the years there is data available. Only this case is studied for all 3 case studies. This represents an overview of what has been happening across time.

To investigate the relation between parameters, two types of analysis were produced. The first one being a simple correlation metric (Pearson) that allows the quantification of how related two different parameters are. In other words, it shows how and how much one parameter is associated or correlated with the others.

A further investigation using SOMs (Self-Organizing Maps) also helps to visually understand how the association between parameters occurs. This kind of Artificial Neural Network (ANN) is often resorted to show correlation between water parameters that are not easily recognizable. They are used because they permit a clear representation of the correlation between parameters of the different sample locations in a 2D plan.

Afterwards, a Principal Component Analysis (PCA) is done. In this part, among other things, the correlation of the different components with the parameter is achieved.

Then, using the once produced SOMs, a cluster analysis is produced. This inspection is relevant, since it allows us to detect sample locations with similar parameters values. With this technique, one is able to identify directly which locations are in the same cluster and inspect why they are put together, thus bringing new knowledge and new questions into discussion.

Following this analysis, several Water Quality Indexes (WQIs) were put in place to analyze water quality in the different networks.

3.2 Artificial Neural Networks (ANNs)

The Artificial Neural Networks (ANNs), already mentioned in this work, are a type of neural network, such as the Recurrent and Convolutional Neural Networks. In an attempt to replicate the human brain, scientists came up with this method. In this kind of models used in Deep Learning, there are three or more layers of nodes (neurons). The first layer, also called input layer, receives the information given by the user. The last layer, also called the output layer presents the final result obtained from the learning

process. In between these layers, there is one or more hidden layers. The neurons in one layer can only connect to the layer immediately before and after. These connections between nodes in different layers have a weight associated to them, which is essential for the training of the model. A node can have multiple input and output links and the output of each node is given by an activation function. A visual representation of a neural network can be found in Figure 3.1.

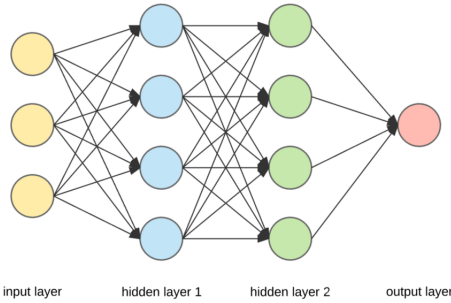

Figure 3.1: Example of a simple neural network with two hidden layers. Figure taken from [1]

The activation function will determine if the neuron is "activated", meaning it will decide if there is any output from that neuron, and if it does, how much of a value it has. This activation happens if the output value is superior than a threshold value defined *a priori*. Employing activation functions in a neural network allows the introduction of non-linear properties in the neural network. They might take the shape of a step function taking values ranging from 0 to 1. Some of the most used activation functions used are:

1. Binary Step or Heavyside Function:

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

2. Sigmoid/Logistic Function:

$$f(x) = \frac{1}{1 + \exp^{-x}} \tag{3.1}$$

3. Hyperbolic Tangent Function

$$f(x) = \frac{\exp^{x} - \exp^{-x}}{\exp^{x} + \exp^{-x}} = 2 \times sigmoid(2x) - 1 \tag{3.2}$$

4. Linear Function

$$f(x) = ax \tag{3.3}$$

5. ReLU Function

$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

A node receives all its weighted inputs and then applies the activation function. Let z be the output of a node as seen in Figure 3.2. It is given by:

$$z = f(x \cdot w + b) = f\left(\sum_{i=1}^{N} x_i w_i + b\right), \tag{3.4}$$

where the product of $x_i w_i$ represents the weighted input i, N is the total number of inputs of the node and b is the bias. The bias is an input to all the nodes and allows to shift the result of the activation function to the left or right. It can also help the training process when all the input variables are 0.

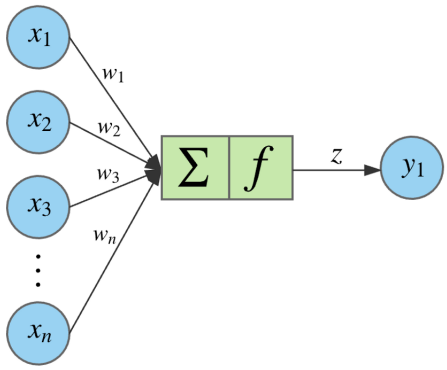

Figure 3.2: Node in the neural network in a forward pass. Figure taken from [1]

Before the learning process, there can be defined hyperparameters (constant values inserted by the user for the training of the model) such as the learning rate and the number of hidden layers. The learning rate defines the size of the change on the model in response to the estimated error each time the model weights are updated. It can be considered a key factor in the training process, because, in one hand, a small value may result in a long training process that could get stuck. In the other hand, a value too large may result in learning a sub-optimal set of weights too fast or an unstable training process. One alternative to solve this kind of issue are adaptive learning rates.

The input presented for the neural network to learn can have an output variable (supervised learning) or not (unsupervised learning) or even additional information given while the training process occurs (reinforcement learning). In supervised learning, the model makes a prediction and then compares its result to what was expected and adapts the model weights to improve the results. The aim of the model is to minimize this cost function. However, in the unsupervised learning the model does not have any information regarding the label or output variable, there are only desired outputs. In this case, the cost function is dependent on the task at hands. In this scenario, it is possible to find a subtype of learning called competitive learning. This form of learning is important for the following section. In this case, output units are said to be in competition for input patterns. During training, the output unit that provides the highest activation to a given input pattern is declared the the winner and is moved closer

to the input pattern, whereas the rest of the neurons are left unchanged. This strategy can also be called winner-take-all, because only the winning neuron is updated. Lastly but not least important, is the fact that output neurons may have inhibitory connections. This means that a winning neuron can inhibit/affect other neurons by an amount proportional to its activation level/value - this will correspond to the neighbouring function in SOMs seen in the next section.

In the third possible case, the reinforcement learning, the output variables are given as input (same as supervised learning), but here the network is also provided with additional information during the training process. Usually what happens is that once the network has calculated the outputs, we are given information about the accuracy of that result, if it is correct or not, and possibly the nature of the mistake that the network made.

There are big advantages while using neural networks instead of the traditional machine learning methods. They have a high performance, solve problems humans may not be able to conceptualize due to the complexity of the network, can be used with regression or classification problems and can handle large amounts of data. Nevertheless, there are some cons associated to this kind of method. It has a ”black box” nature, meaning it is hard to comprehend all the numerical steps in between the input and output layers and consequently it can be hard to understand why the neural network makes some conclusions. They also take longer and require more data than some other classical methods to train and demand a significantly high computational power.

3.3 Self-Organizing Maps (SOMs)

In the interest of showing a clearer representation of the data, Self Organizing maps or Kohonen maps are used to reduce dimensionality and represent data in a 2 dimensional lattice for a more simplified and easy to read map. It is an artificial neural network that is trained using unsupervised learning, because there is no output variable to produce the desired 2D map. Instead of using error-correction learning, it uses competitive learning as it will be explained further in this section. Introduced by the Finish professor Teuvo Kohonen in the 1980s, it is very used in this kind of studies, due to its ability to represent complex data in a very simplified way, since the input vectors with common characteristics are assigned to the same or neighbouring neurons and thus preserving the topographic properties of the original data.

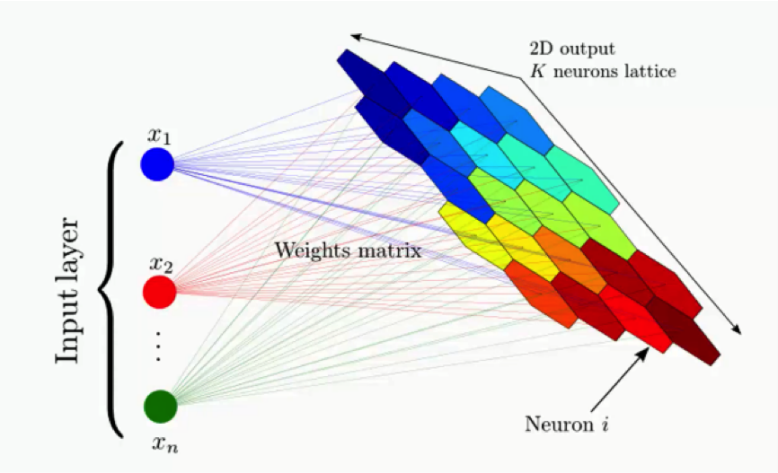

Figure 3.3: Structure of a SOM for a K number of output neurons. Figure taken from [2]

This procedure takes an input layer that is connected to each vector of the data set. In other words, every node of the input vector corresponds to a different variable. This layer is then linked to an output/competitive layer made of an array of nodes with weights associated to them. The length of this weight vector is equal to the number of nodes in the input layer and is given an initial value. Then, the training process initiates and the output node with the closest distance to the data set point is activated and is denoted the Best-Matching Unit (BMU), hence the competitive learning. The nearest neighbours of the activated neuron are also activated according to a neighbourhood function. This neighbourhood function range of action will decrease over time, meaning that in the initial steps it will have an impact in a large number of neighbouring nodes, and by the final steps, each time a new BMU is chosen, this neighbouring function will only effect a small amount of nodes. One neighbourhood function typically chosen (used in this thesis) is the Gaussian function. The weights are then updated and we proceed to the next observation and do the same process to get a new BMU and new updated vectors. Each observation is a vector $x_1, x_2..., x_L$ of dimension L. The algorithm is as follows:

1. Initialization: Set the initial weight vectors in the interval $[0, 1]$. There is a vector for each output node with a dimension corresponding to the observation dimension. This can be seen as a weight matrix of elements w_{ij}, $i = 1, ..., S$ and $j = 1, ..., L$, where S is the number of output nodes in the output layer and L the number of evaluated parameters. The initial learning rate $\eta \in]0, 1[$, the map size, the neighbourhood radius or neighbor ratio R and the number of maximum iterations are also defined.

2. Distance calculations. Select an input vector $x^k = (x_1^k, x_2^k, ..., x_L^k)$, where $k = 1, ..., M$ with M being the sample/observation number. The distance of the vector to the weight vector is then calculated

using a distance measure, in this case, the Euclidean. This distance is calculated as:

$$d_i = \sqrt{\sum_{j=1}^{N}(x_j^k - w_{ij})^2}, i = 1, ..., S \tag{3.5}$$

3. Selection of the BMU. Perceive the node with the smallest distance - this node is called BMU.

4. Update. The weight vector w_{ij} and the neighbourhood radius are updated.

$$w_{ij}(t+1) = w_{ij}(t) + R(t)\eta(t)(x_j^k - w_{ij}(t)) \tag{3.6}$$

where $w_{ij}(t+1)$ is the weight vector at time step $t+1$. The learning rate $\eta(t)$ and the neighbourhood radius $R(t)$ depend on time t, since they decrease with the number of iterations.

5. Recursion. Since the SOMs are an iterative process, the method continues until the maximum number of iterations is reached, and then go back to point 2.

Regarding the map size, from [9] and [43] the optimal SOM size defined is $5\sqrt{N}$, where N is the total number of observations in the data set. This optimal size will reduce the number of too many empty cells in the map, giving a more accurate representation.

3.4 Other Unsupervised Methods

3.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique mainly used to reduce the dimensionality of a data set. This method creates new uncorrelated variables that aim to maximize the variability of the original variables. These new variables are called principal components (PCs) and are calculated solving an eigenvalue/eigenvector problem. This ends up being an eigenvalue problem in the sense that the principal components are the eigenvalue/eigenvectors of the covariance or correlation matrix. Since the values of the sample principal components depend on the data set at hand, PCA is considered an adaptive data analysis technique.

Each principal component is formed as a linear combination of the original variables in the data set that maximizes the variation/information in the data. Another great feature of this method, is the fact that each new principal component is orthogonal to all the previous ones already calculated. Since the original variables can be replaced by the principal components, if one decides to create these new variables and utilize them, their meaning will be compromised.

There are several ways to choose how many principal components to keep. Two of the most used are essentially manually picking how many dimensions to keep for the data set. The number of dimensions is equivalent to how many principal components one chooses to maintain. Since this is not the desired objective, one other way is to select the number of principal components and define a percentage threshold of explained variance. For example, setting a threshold of 90%, the number of components picked is the number of components that, summed up, total or top this explained variance percentage.

3.4.2 K-Means

K-Means clustering is one of the simplest unsupervised machine learning algorithms. *A priori* defining a target number of clusters k, this algorithm identifies k centroids and then allocate each data point to one of the centroids in a way that minimizes the distances between points and centroids.

The optimal number of clusters for this method can be given by the elbow plot. For an increasing number of clusters an error function is calculated. This error function can be the sum of squared distances from each point to its assigned center (distortions). In this way, the optimal number of clusters is given by the number that is located on the elbow of the generated plot of the error in function of the number of clusters. Further information on this algorithm can be found in [44] and [45].

3.4.3 Hierarchical Clustering

Like K-Means, hierarchical clustering is an unsupervised learning algorithm that aims to group the data points in clusters. This method to cluster data points can be agglomerative, where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy, or divisive, where all observations start in one cluster and splits are performed recursively as one moves down the hierarchy. Unlike the K-Means algorithm, the number of clusters k is not specified. Hierarchical clustering has a visual representation through a dendrogram. A dendrogram is a cluster tree diagram, where the distance of merge of data points/clusters is recorded. The agglomerative hierarchical clustering will be the technique used.

In this work, the method used to calculate the distance between clusters is the Ward's minimum variance method and the distance considered was the squared euclidean distance (3.7).

$$d_{ij} = \|x_i - x_j\|^2 \tag{3.7}$$

where d_{ij} is the squared euclidean distance between points x_i and x_j with $i \neq j$.

The optimal number of clusters for this technique can be analyzed using the dendrogram. The largest vertical distance without any horizontal lines is identified and a line is drawn in that zone. The number of vertical lines crossed by the drawn line is assumed to be the optimal number of clusters [46] [47] [48].

4

Data Pre processing

Contents

4.1 Parametric values

The variables kept for each data set are the name of the sampling site identified by an unique *objectid* number or its actual name, the date of when the sample was collected (that was used to organize the data by year to see how the parameters evolved throughout time that observations were gathered), the parameter evaluated at that specific place and day and the unit used to measure it.

Since not all the parameters are of interest, a choice had to be made regarding which of the parameters are significant to inspect. Hence, the only ones picked are the ones present in *Diário da República*, which determines the parametric values for some important microbiological, chemical and indicator parameters in a water supply network. These values can be found in Decreto-Lei n.° 306/2007, of August 27th, with the changes introduced by Decreto-Lei n.° 152/2017, of December 7th. These parameters and their corresponding parametric values present in the data sets are in the following Tables: the parametric values and units for the microbiological parameters can be found in Table A.1. In Table A.2 for the chemical, and in Table A.3 for the indicative ones.

The values of the parameters are positive continuous variables that vary in magnitude according to what parameter is being studied and what unit is being used to measure it.

4.2 Regularity of Analysis

Some of the parameters are considered more important to monitor than others for the most diverse reasons. Among these causes, health and quality assurance rank the highest. Some water quality control parameters are monitored more often than others, according to the Portuguese law (DL 152/2017), which establishes three categories for sampling frequency. Ranked from the most to the least important, the 3 categories are Routine Control 1, Routine Control 2 and Inspection Control. In the first one, there are only 3 parameters: *Escherichia coli* (E. coli), coliform bacteria and residual disinfectant. In the second category, there can be found the parameters: smell, taste, pH, conductivity, turbidity, enterococci and both the number of colonies at 22℃ and 37℃. The parameters not mentioned in these two categories fall into the 3rd category, where the amount of analysis required is the least. There are some exceptions regarding the parameters that can move from the Routine Control 2 to the Inspection Control to the Routine Control 2 that will not be mentioned here as they can be found in the Decreto-Lei already mentioned.

One important feature of these categories is that the analysis of a less important category requires the analysis of all the more important ones. In other words, for example, if one entity has to analyze the parameters present in the Routine Control 2 they will also have to test the parameters in the Routine Control 1. In the same way, if they have to inspect the values of the parameters for the Inspection Control, they are automatically obliged to check out the values for the parameters in both the Routine

Control 1 and 2.

4.3 Parameter Selection

Only a few of the parameters already mentioned were selected to be included in some part of the upcoming analysis, because not all of them have the same relevance. Another reason for exclusion can be an insufficient number of observations. Another possibility is if there is just not enough different values for the parameter in question. In this sense, for all the 3 case studies the number of parameters investigated was reduced once again.

After some considerate analysis of the problem and the different data sets, it was decided to keep, if possible, 14 different parameters that would allow, up to some extent, evaluate the water quality in the different distribution networks. These 14 are the number of colonies at 22℃ and 37℃, conductivity, hardness, iron, manganese, nitrates, oxidability, pH, residual disinfectant, temperature, THMs, Total Organic Carbon (TOC) and turbidity. Notice that if a water utility does not have all 14, then it will utilize as many as possible.

4.4 Process the data

Upon looking at the data set received from the water utilities, some changes had to be made. These changes were essential, since it allowed the data set to be properly organized to be able to do the following analysis.

Not all of the columns in the data set received from the water utilities were of interest and so, some of them were immediately discarded and only the relevant ones were kept. Since there were several equal parameters written in different ways, one had to be consistent and name them the same way to make sure that the same parameters were not being treated differently. An *objectid* attribute was also added. This *objectid* is unique ID for each sampling site. Regarding the actual values of the parameters, a consistency check was also done. Since not all values were actually numbers, they had to be converted. Some of these values came with an associated mathematical symbol ($<$ or $>$) and in order to deal with them, it was decided to change the value according to the sign related to it. According to the scale of the results of the parameter, those values were slightly lessened or increased.

Another change was the addition of some other parameters. Some of the parameters in the DL 152/2017, are a combination of different parameters and upon looking at the data set received, it was noticed that some of these parameters were not present, but only their ”components”. So, from these ”smaller” parameters, other parameters were built and integrated in the data set. One example, is the Trihalomethanes (THMs) that are actually the sum of the concentrations of 4 parameters: chloroform,

bromoform, bromodichloromethane and dibromochloromethane. In the other way around, the initial ”components” were removed from the data set, as well as some observations of parameters feaured in DL 152/2017.

The data set initially utilized contained only the 14 most important parameters as explained in Section 4.3. The next step, after this first changes to the data set, was to investigate outliers. One observation (the pair sampling site and date) was considered an outlier if, for one parameter, its result is completely different from what is expected. This inspection was done manually. The removal of one outlier for a certain parameter resulted in the elimination of this only value, not invalidating all the other parameters evaluated in that observation.

These nonsense values may have origin in a defective equipment, in human error (for example, a typo when introducing the values in digital format) among other causes. This detection was done by manually investigating the results obtained for the different parameters and removing entries that did not seem to be fair.

One example of this is the outlier found for the parameter THMs for the Barreiro data set seen in Figure 4.1.

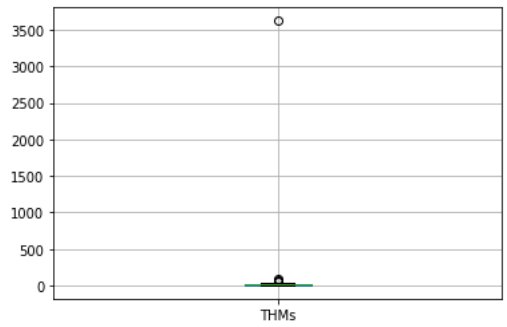

Figure 4.1: Boxplot with outlier identification of the THMs parameter in the data set.

As to the actual values found for the selected parameters an evaluation of its statistical characteristics was conducted.

5

Methodology

Contents

5.1 Correlation

With the processed data, for each case studied, a Pearson correlation matrix was built and the important values are highlighted in a heatmap. Observing this map, some conclusions are drawn. Any correlation above the absolute value of 0.6 was considered of interest and was the object of a further investigation. In this examination, the focus is to demonstrate how the the evolution of the correlation between pairs of parameters occurs.

5.2 SOMs

Regarding the application of these maps to the actual data provided, the data sets utilized in this section are the ones used to calculate the correlations.

In order to initialize the SOM, a map size has to be given. As seen before, the ideal number of nodes in the output layer corresponding to the number of cells in the 2D map is close to $5\sqrt{N}$, where N is the total number of observations. The procedure used was to define the optimal size of the map and then train 10 different models and retrieve their respective topographic and quantization error. Then, from the 10, select the one with the lowest topographical error. This error was the one chosen, since the quantization one doesn't show much variation between models.

These errors were used to quantify the quality of the maps and can be seen as quality measures. The quantization error is computed by summing the distances between the nodes and the data points. In other words, after the winner-take-all learning process in the SOM, each image input vector x_i becomes associated to its best matching model on the map $\phi(x_i)$ (the operator ϕ is the mapping from the input space to the SOM). The quantization error is nothing more than the measure of how close the final SOM is to the original input value and is given by:

$$QE = \frac{1}{N}\sum_{i=1}^{N} \|\phi(x_i) - x_i\| \tag{5.1}$$

where n is the total number of observations. The topographic error accounts for a SOMs preservation of local topological features in a low dimensional output space. In other words, the topographic error is a measure of how well the structure of the input space is modeled by the map. It is calculated by finding the best-matching and the second-best-matching neuron in the map for each input and then evaluating the positions. If these two nodes are next to each other, then it's fair to say the topology has been preserved; if not, it is accounted as an error. The topographic error is given by the total number of errors divided by the total number of observations n. So, let $\mu_1(x)$ and $\mu_2(x)$ return the best-matching and second-best-matching neuron for a data point x and $t(x)$ be the error function that returns 0 if $\mu_1(x)$ and

$\mu_2(x)$ are neighbours and 1 otherwise, the topographic error can be expressed as:

$$TE = \frac{1}{N}\sum_{i=1}^{N} t(x_i) \tag{5.2}$$

Since the data sets do not have values for every parameter for every location and date, there was a problem regarding the filling of some missing values. Because some variables had a big portion of missing values, it would not make sense to fill those missing values with some metric, since it would wrongly represent the variable in question. To handle this problem, the average was applied to all the missing values if a parameter/variable has more than 66% values. In other words, if a variable has more than 34% of missing values it is removed from the model; for the remainder of the parameters the average is applied to those missing values.

So, for every SOM map produced, it is shown the corresponding parameters that made the cut, the total number of observations used as input to train the SOM, the optimal number of nodes used and the map dimensions. The map dimensions are calculated to be as close as the total number of nodes. The 10 trained models errors for each SOM will not be displayed and only the errors of the best model are revealed.

Furthermore, a hits map can be produced. The topological structure of this map is the same as a component of the components plane. It represents the number of times each map unit, or neuron, was the BMU for each input register, so that the distribution of the BMU for a given data is represented. This gives an idea of the number of input observations that gather in each neuron. This makes possible to compare the importance of each unit in the components plane. This kind of map is in the Appendix section since it only serves as a supplement for the components map. The SOMPY [49] library was used to produce the maps in Python. It is an adaptation from the `somtoolbox` from Matlab.

5.3 Principal Component Analysis (PCA)

In this thesis, the PCA is mainly used to see and analyze how the principal components correlate with the variables of the original data set, which correspond to the water parameters. These correlations values are the coefficients of the different parameters of the principal components. Regarding the application of this analysis, a set number of principal components is defined to achieve a reasonable percentage of variability explained. This value was considered acceptable around 85-90% or above. Then, the correlations between the different principal components and the variables/parameters were analyzed. Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either direction. In this thesis, a correlation above the absolute value 0.5 is considered important.

5.4 Cluster Analysis

For this study, a K-Means and the described hierarchical clustering method were applied. To obtain the optimal number of clusters for the K-Means technique, the elbow method/plot was used. As to the hierarchical clustering, the dendrogram was the followed option. Usually, there was a concordance of both methods in the optimal number of clusters found.

In order to perform the analysis of the clusters, a data normalization was done. This is achieved by subtracting the mean $\bar{x}$ and dividing by the standard deviation s for each parameter/variable. In a mathematical representation, for each entry x, the normalized observation $\tilde{x}$ is given by:

$$\tilde{x} = \frac{x - \bar{x}}{\mathrm{s}} \tag{5.3}$$

This allows to transform the data from any scale to a smaller range. It is helpful to have a better representation of the data when the objective is to compare different parameters at once. After that, an individual analysis of each cluster is accomplished. to interpret and determine if the patterns found have impact and meaningful.

The results obtained for each cluster are then represented in the SOMs for a clear idea how the different clusters are distributed and how the values associated to each parameter relate to the clusters.

The data served as input to produce the cluster analysis is the same data used to recreate the SOMs. The clusters used to produce the further analysis derive from the K-Means technique. This was an arbitrary choice, since both clustering methods show similar results.

5.5 New proposals for Water Quality Indexes (WQI)

From the previously discussed Water Quality Indexes stated in the State of the art section, new indexes were created. The objectives defined for this part of the thesis is to show in a global view how the different indexes can or can not be representative of the quality of the water. Another important factor regarding the indexes chosen/created was to include as much information provided by the data as possible. The indexes based on rating curves were not considered. This is due to the fact that these rating curves are often given by experts in the water quality analysis and/or experts of the specific region where the studies take part. So, with this in mind, 3 different indexes A, B and C were designed using mixing ideas and algorithms from previous works.

5.5.1 Index A

The first index analyzed, from now on mentioned as index A, is very similar to the previously discussed index 1 and 2 seen in Table 2.1, and applied in a similar fashion as Alomran et al. (index 4). [28]. The index is built as follows:

- Quality index (Qi): $\frac{C_i - C_0}{S_i - C_0} \times 100$
- Weight index (Wi): $\frac{1}{S_i}$
- WQI: $\frac{\sum^N W_i Q_i}{\sum^N W_i}$

where C_i is the concentration for the parameter i, C_0 is the ideal value for the parameters (7 for the pH and 0 for all the others), S_i is the standard value for the parameter i and N is the total number of observations.

Regarding the scale to quantify the actual quality of the water it is considered the same used in works such as Akter et al. [20] and Yisa et al. [19], since the WQI used is the very similar. This is represented in Table 5.1.

Table 5.1: WQI classification for index A and B

WQI Range	Water Quality
$<$50	Excellent
50-100	Good
100-200	Poor
200-300	Very Poor
$>$300	Unsuitable for drinking

In order to deal with the microbiological parameters, such as *Escherichia coli* (E. coli) and coliform bacteria - that have a $S_i = 0$, each observation was analyzed and a flag was created. If the value for these parameters is 0, then the analysis of the remaining parameters is done. If one observation has a value over 0 for at least one of the parameters, it is assumed that the water is contaminated and so undrinkable for human consumption. For this measure, a final WQI with a value of 301 was assigned to such observations. For this index, the residual disinfectant and hardness are not considered, since they don't have an associated S_i nor C_0 value, but instead have an interval.

5.5.2 Index B

The second index proposed, index B, takes inspiration from index 3 from Table 2.1 and is utilized in the same way as index A, with the creation of flags for microbiological contaminated water. Despite being built similarly to index A, the modification was to use a different weight index for each parameter. The weights were defined according to the number of times they were evaluated, meaning that a parameter

that is tested more often is considered more important than another one not tested so often. This represents the only difference from the first index. The flag system is the same as in index A. This assignment of weights according to the importance of a parameter is similar to the index 3 mentioned in the state of the art section in Table 2.1. Considering N observations, this index is built in the following way:

- Quality index (Qi): $\frac{C_i - C_0}{S_i - C_0} \times 100$
- Weight index (Wi): $\frac{w_i^*}{\sum^N w_i^*}$
- WQI: $\sum^N W_i Q_i$

The weights of the parameters are as follows:

- w_i^* = 3: Coliform bacteria, residual disinfectant and *escherchia coli* (E.coli)
- w_i^* = 2: amonnium, smell at 25℃, conductivity, color, manganese, nitrates, number of colonies at 22℃ and 37℃, oxidability, taste at 25℃, turbidity, pH,
- w_i^* = 1: 1,2 - dichloroethane, aluminium, antimonnium, arsenium, benzene, benzo(a)pyrene, borum, bromates, lead, cyanides, chlorides, clostridium perfrigens, copper, chromium, cadmium, calcium, indicative dose, hardness, enterococci , ethenes, iron, fluorites, magnesium, nitrites, nickel, Polynuclear Aromatic Hydrocarbons (PAHs), radon, selenium, sulfates, sodium, THMs.

These weights reflect the importance of the parameters, where a weight of 3 reflects a parameter of the most value and a weight of 1 reveals a parameter that is not that important. The parameters were introduced accordingly to the number of observations present in the data set. This count is in conformity to the three categories found in DL 152/2017. In other words, a weight of 3 corresponds to the parameters in Routine Control 1, a weight of 2 is where the parameters in Routine Control 2 are inserted and finally, the other parameters in Inspection Control have a weight of 1.

The same classification as index A is used for this index (Table 5.1).

For these two indexes, the overall WQI of the network is given by the mean of the index value calculated for every observation.

5.5.3 Index C

The third index developed or index C is a new proposal not bored in the previous ones. Here, a method of differences was applied to deal with all the parameters that have a standard value that consists of a range of values instead of a single standard value. In this way all the information provided is used. If a parameter has a unique standard value and not a range of values, it is assumed that the lowest that value is, the better the quality of the water. Apart from the pH that has an ideal value (and also

an interval of parametric values), the residual disinfectant and hardness don't have this kind of value and are yet to be utilized to characterize the quality of the water in a network. The method starts by normalizing the data, so that all parameters have the same scale. The objective of this index is to sum the differences of the different concentrations of the parameters to the respective parametric values. In a more practical case, there can be two different instances. The first and most simple one is when the standard value is a single value. Here, for each parameter, the difference between the standard value S_i and the concentration C_i is calculated, added as a new variable and stored. The second case is when the parameter has a range of optimal values, such as the pH, residual disinfectant and hardness, also normalized. Three possible outcomes can happen in this scenario:

1. If the concentration of an observation is within the interval of values, the difference to both ends is summed;

2. If the concentration is over the top limit, then the difference between the top limit and the concentration is calculated. The same reasoning is done if the concentration is below the bottom limit;

3. The third is related to the signal of the differences measured, i.e. if a concentration is below the standard value or within the range of values, then the difference is positive and is viewed as a good thing. If the concentration is over the standard value or outside the range of values, then the difference calculated will always be negative, and is then viewed as a bad indicator.

This index allows to penalize a parameter if, for a given observation, its concentration is over the parametric value in a proportional way of how far (bad) it is from the actual limit. In other words, if a registered value is very off the boundary limit, it will have much worse consequences and a consequent worse index than one other observation that is just barely over the limit. The same thing applies for the good values. If for a parameter, an observation has a value very close to 0, then it will have a better index than another observation with a value very close to the stipulated limit.

For the microbiological parameters (E.coli, coliform bacteria and enterococci), it was also implemented a flag method that detected if one observation had a superior value of 0, for any of these parameters. If that was the case, an index of -30 was attributed to it automatically. This value classifies the water as undrinkable, according to the classification being used. This value was the one used, because while assigning the water as undrinkable, this value adjusted the overall WQI of the network to be reasonable and show appropriate results.

Another last consideration, is the fact that for the three parameters that have a range of parametric values (pH, hardness and residual disinfectant) the values that are within the interval (considered good results) all have the same distance (positive value). This means that for this case, one can't actually measure how good the result is. On the other hand, if a value is outside that range, it is possible to

quantify how bad that value actually is and therefore penalize according to the difference measured.

To keep the scale of the difference of the parameters, the values (actual concentrations and standard values) were normalized using the min-max feature scaling. This is an important step, because the different parameters are measured in different units, and so, if one was to calculate the differences of the concentrations of the parameters and their respective standard values, there would be massive differences. The min-max normalization is as follows:

$$\tilde{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{5.4}$$

where $\tilde{x}$ is the rescaled value, x the original value and $min(x)$ and $max(x)$ the corresponding minimum and maximum values of the parameter. This normalization technique was preferred over the standard deviation scaling, because there were some parameters that contained a lot of zeros, including their respective parametric value and so dividing by the standard deviation did not make much sense.

In order to stay truthful to the amount of measures some observations have - one observation is the pair sampling point/date - the final sum of the differences is multiplied by 100 (just for scale measures) and is divided by how many parameters are being analyzed in the corresponding observation. This is done, because one observation can have a lot of parameters under the parametric value or within the range of appropriate values, resulting in an important sum of differences and thus a very high value for the index. Meanwhile, one observation that has few parameters being analyzed and all in good conditions as well, would have a very small sum of positive differences. And so, even though both observations have a very good water quality, the first one would possess a very high index compared to the second one, just because it has more parameters being tested. The dividing process allows a certain ”normalization” regarding the number of parameters being involved in the sum of differences.

To sum up this index, it can be explained in the following steps:

1. Normalize the data and parametric values (values and ranges);

2. For each parameter i, calculate the difference of the concentration C_i to the corresponding parametric value. Here, there are 2 cases to consider:

   - Case 1: Parametric value is a number S_i. In this case, Calculate the difference $d_i = S_i - C_i$;
   - Case 2: Parametric value is a range of numbers. Here, if the value is within the two numbers $[a, b]$, then the difference is the sum of the differences of the concentration to both ends of the interval $d_i = (C_i - a) + (b - C_i)$. If $C_i > b$, then $d_i = b - C_i$. Lastly if $C_i < a$, then $d_i = C_i - a$.

3. Sum all the calculated differences for each observation: $d_2 = \sum^N d_i$, where N is the total number of observations;

4. Calculate how many parameters are being analyzed for each observation. Let this sum for each observation j be denoted by P_j, $j = 1, 2, ..., N$;

5. $WQI_j = \frac{d_2 \times 100}{P_j}$, $j = 1, 2, ..., N$.

The final WQI is given by the mean of all WQI_j, or the water quality index of all the observations.

Regarding the quality of the the values obtained for this index, they can be classified in several aspects. These ratings go from excellent water to unsuitable for drinking. The classification of the results obtained can be seen in Table 5.2.

Table 5.2: WQI classification for index C

WQI Range	Water Quality
$<$-10	Unsuitable for drinking
-10-0	Poor
0-25	Good
25-50	Very Good
$>$50	Excellent

These intervals were the ones chosen with the respective ratings, because they allowed a separation of different instances for all the case studies. They were defined after being measured for the cases in this thesis. This kind of ratings allows further and direct conclusions. The water quality defined as good is the case when only the 3 parameters in the Routine Control 1 (coliform bacteria, E.coli and residual disinfectant) are within standard/parametric values. The other positive ratings mostly depend on the the number of tested parameters for every observation. Water qualified as ”very good” is usually the case when several parameters, more than the 3 already mentioned but not all of them, are tested and the vast majority is within legal values. The ”excellent” rating is typically attributed to observations with almost all the possible parameters present in the data set evaluated and all of them (or a very big proportion) are within parametric values. In the other way around, a ”poor” water is generally the case when the three parameters in the Routine Control 1 are the only being tested and only the residual disinfectant is not within the standard values. Observations where the majority of parameters being tested fail to be in the parametric values or observations where one parameter is very off the parametric value can also fall into this category. Water ”unsuitable for drinking” represents the extreme cases of these last two possibilities and also the presence of microbiological activity, such as E.Coli and/or coliform bacteria.

Unlike the other indexes (A and B) there is not a ”very poor” rating, simply because the range of actual values of the WQI for the different observations does not justify a new class of classification.

6

Case Studies

Contents

6.1 Barreiro

6.1.1 Statistical Analysis

The statistical characteristics of the selected parameters of the Barreiro data set can be found in Table 6.1 after the outlier removal. There are a total of 2192 observations with 47 different parameters evaluated. The main descriptive statistics were calculated for each of the parameters in all the years that there are observations. In this description, it's included the sample mean ($\bar{x}$), standard deviation (s), the minimum and maximum values, the parametric value (PV) of the parameter and the total number of observations with an actual value.

Table 6.1: Descriptive analysis of the parameters - Barreiro overall view

Parameter	$\bar{x}$	s	Min	Max	PV	no. Obs.
Colonies at 22℃*	17.71	56.80	0.0	301.0	100	783
Colonies at 37℃*	20.62	60.15	0.0	301.0	20	783
Conductivity	334.03	127.27	108.0	2170.0	2500	783
Hardness*	122.87	61.14	17.0	490.0	150-500	128
Iron	48.59	30.38	20.0	320.0	200	128
Manganese	13.14	4.58	5.0	46.0	50	721
Nitrates	8.65	4.99	1.0	100.0	50	723
Oxidability	0.96	0.29	0.60	4.60	5	721
pH	7.46	0.39	6.10	8.50	6.5-9.5	783
Residual Disinfectant*	0.37	0.18	0.10	1.50	0.2-0.6	2188
THMs	13.28	13.87	0.70	96.0	100	128
Turbidity	0.55	0.41	0.40	6.20	4	783

Parameters noted with * are considered only recommended and not mandatory to comply with the legislation. Overall, all parameters show reasonable values and the majority of them are within the parametric values. Inspecting more closely, it can be seen that parameters such as the number of colonies at 22℃ and 37℃ and the residual disinfectant have more observations outside their respective limit values. Regarding the hardness, its values seem a bit off and there is a large quantity of them that does not respect the parametric value. One reason can be because these values are only recommended and not mandatory to comply with. The statistical evolution of the parameters can be seen in the Appendix Section.

6.1.2 Correlation Analysis

Regarding the original data, one alteration was done. There was a trimming of the data regarding the parameter residual disinfectant. Since it is evaluated/tested way more often than the rest of the parameters, one option took was to eliminate all rows that only contain information regarding this parameter. In practical terms, this means that every observation has at least 2 parameters being tested. Despite

losing some information regarding this parameter (since there is a removal of info), this change is helpful to find the correlations between this parameter and the remaining and remove a lot of missing values from the data set. After this inspection, this filtered data set is trimmed to 780 observations with the selected 12 parameters.

The respective heatmap of the filtered data set is found at Figure 6.1.

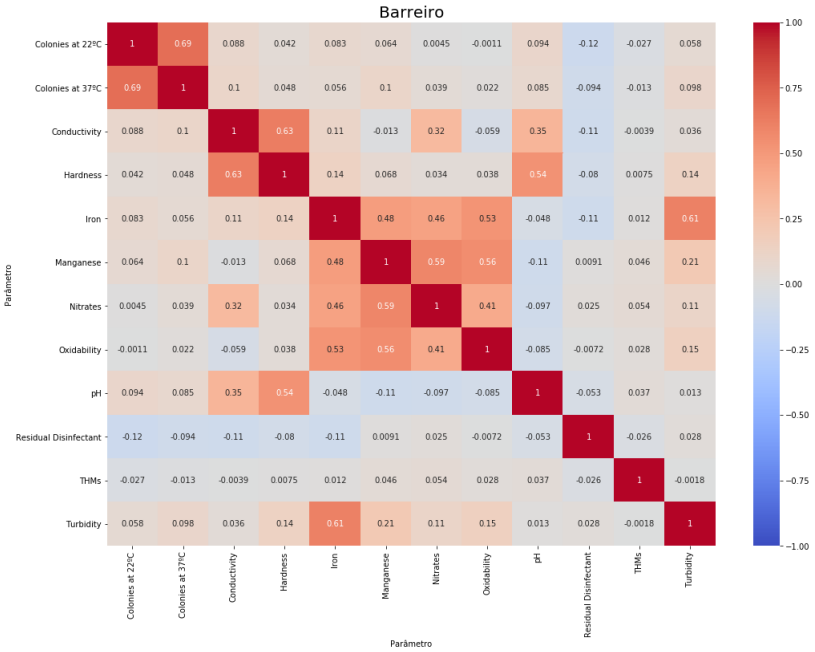

Figure 6.1: Correlation Matrix of the Barreiro Overview filtered data set

With this data set containing almost all available and relevant information, it should be more clear which parameters correlate with each other and how strong is that link.

The association between the number of colonies at 22 ºC and 37 ºC seems evident and has a value of 0.69. This value of correlation indicates that when microorganisms are found in Barreiro drinking water, those microorganism frequently include species able to grow at 37 ºC, i.e., able to infect humans.

One other pair with a significant value of correlation is the pair hardness and conductivity with a value of 0.63. Being the hardness of water the sum of calcium and magnesium ions, the high correlation with the conductivity indicates that the ability of water to conduct electricity is very much due to the referred ions.

The last one that could be of importance is the link between the turbidity and and iron, with a value of 0.61. Even though there was no information regarding the iron before 2013, this relation seems to be quite significant.

A closer look at how these pairs of parameters evolve throughout time is illustrated in Figure 6.2.

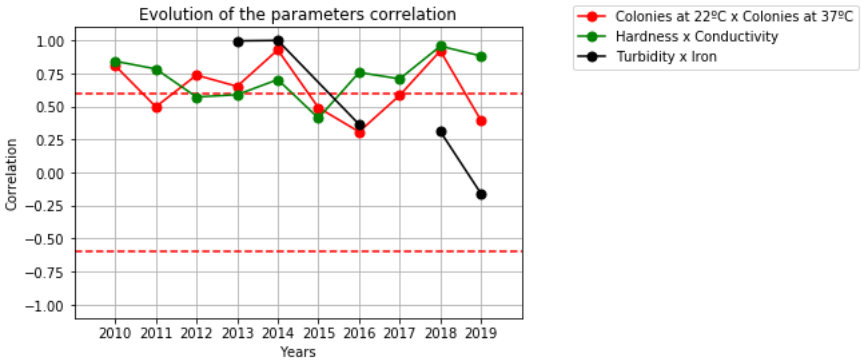

Figure 6.2: Correlation of parameters throughout time in Barreiro Overview filtered data set

From this Figure, it is possible to arrive at several conclusions. Note that if there is not a point at some year, it is because it was not possible to calculate that correlation either because there was not enough variability in the data or simply because that parameter was not analyzed at all that year.

Two pairs of parameters seem to show a very stable high correlation such as the number of colonies at 22℃ and at 37℃ (in red) and the hardness and conductivity (in green). As for the turbidity and iron (in black), their correlation is high in the early years when iron is evaluated. However, for the last 2 years their correlation significantly dropped and an investigation of why this happened could be of interest for the water utility.

To sum up, several correlations are of interest and should be monitored closely to see how they evolve. A further inspection on where on the network these correlations are more meaningful can help the water utility to forecast beforehand where the values will be over the parametric value and the implications to the other parameters.

6.1.3 SOMs

The starting data sets utilized in this section are the ones used to calculate the correlations after the trimming of the residual disinfectant and the non relevant parameters as mentioned in the correlation subsection above.

After analyzing the missing values for each parameter, on one hand, the hardness, iron and THMs were removed. On the other hand, manganese, nitrates and oxidability had their missing values replaced by their respective mean.

The rest of the parameters are used to train the model that produces the map. The best model obtained had a topographic error of 0.0115 and a quantization error of 0.2441.

The total number of observations when grouping all the years is 780. This corresponds to 140 optimal nodes and the chosen map size is 14×10. The respective map is shown in Figure 6.3. The corresponding hits map can be seen in the Appendix Section.

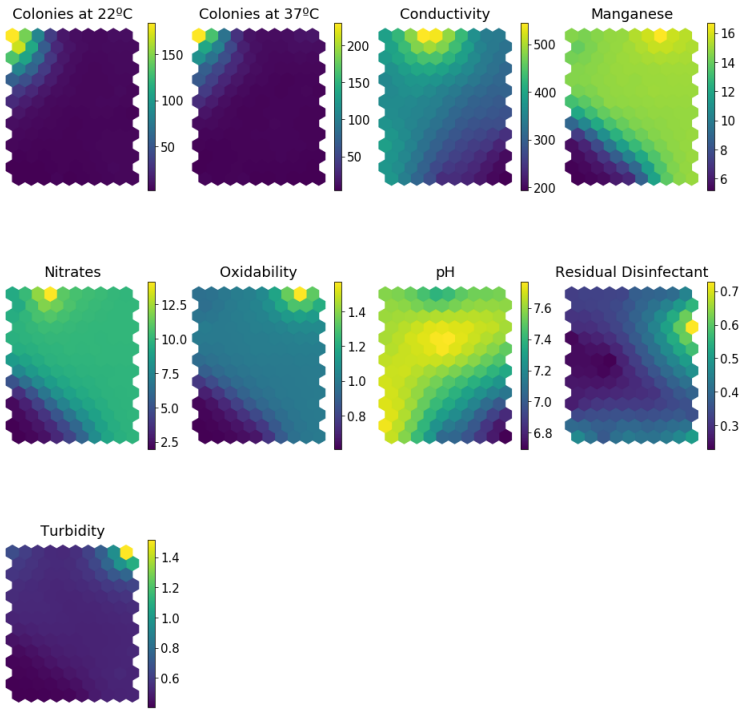

Figure 6.3: SOM of Barreiro Overview filtered data set

Once more, as expected, the number of colonies at 22℃ and 37℃ relate very strongly with each other. There is a significant link between nitrates and oxidability. This suggests that increases in oxidability in Barreiro water are probably a consequence of higher manganese and other ions (e.g., iron) content in the groundwater, rather than due to an increase in organic matter. The colonies at both temperatures correlate in a weaker way with nitrates. Nitrates also correlate in a stronger manner with conductivity. This indicates that by simply measuring conductivity - a parameter that can be measured on site in a reliable way using a cheap probe - one can infer about the nitrates content in Barreiro water. The residual disinfectant has a weak inverse link with both number of colonies which makes sense in a physical way. Surprisingly, the high colonies counts were observed despite the residual disinfectant content was above 0.3 mg/L Turbidity correlates in a feeble way with manganese and oxidability.

6.1.4 PCA

For the data set containing the information regarding all the years, 6 principal components were needed to achieve a total of 88.27% of explained variance. Each component explains 24.04%, 20.0%, 14.64%, 11.09%, 10.13% and 8.38% of the total variance, respectively. In Table 6.2, the correlations of the different principal components with the parameters in the original data are represented. These are often called factor loadings.

Table 6.2: Barreiro overview filtered data set - Principal components analysis

Parameter	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Colonies at 22℃	0.18	0.59	-0.32	-0.01	-0.15	-0.01
Colonies at 37℃	0.21	0.58	-0.31	0.05	-0.14	0.03
Conductivity	0.13	0.28	0.67	0.01	-0.12	0.38
Manganese	0.57	-0.16	-0.05	-0.05	-0.02	-0.15
Nitrates	0.53	-0.13	0.25	-0.09	-0.23	0.32
Oxidability	0.49	-0.21	-0.07	-0.09	0.07	-0.47
pH	-0.05	0.33	0.53	0.22	0.07	-0.67
Residual Disinfectant	-0.03	-0.22	-0.09	0.74	-0.62	-0.04
Turbidity	0.23	0.04	-0.06	0.62	0.70	0.25

Starting with the 1st PC, there is a positive significant correlation with manganese and nitrates. It could be evidence that these 2 parameters are correlated with each other, meaning high values in a site of one parameter might have high results for the other.

The 2nd principal component has high correlations with the number of colonies at 22℃ and 37℃. This is an indicator that this component increases with increasing number of colonies at 22℃ and 37℃. It can be seen as a measure to evaluate the number of colonies present in the data set in this year.

For the 3rd component, it is observable that this component increases when conductivity and pH increase, since this PC is highly correlated with both of these parameters. So, this PC can be seen as a measure of how conductive and alkaline the values can be in the data set.

The 4th principal component relates highly with both the residual disinfectant and turbidity. This means that this PC increases with increasing values for both of these variables. This suggests that places with high values of residual disinfectant also show high results for the turbidity.

As for the 5th PC, it correlates negatively with the residual disinfectant and positively with turbidity. This contradicts the conclusions drawn from the 4th PC, as the signs of the 2 parameters are no longer in concordance. So, this time around this PC increases with the increase of turbidity and decreases with the residual disinfectant parameter.

The 6th and last principal component correlates negatively with the pH. This suggests this PC is a measure of acidic the water is in the sampling sites.

6.1.5 Cluster Analysis

Regarding this kind of analysis, the elbow plot and dendrogram were produced to find the optimal number of clusters to analyze. From them, it was observable that the optimal number of clusters to use was 5.

The representation of these clusters in the already mentioned self organizing maps is observed in Figure 6.4 and 6.5.

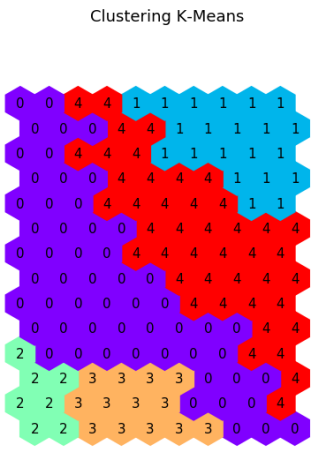

Figure 6.4: Barreiro Overview filtered data set - K-Means in SOM

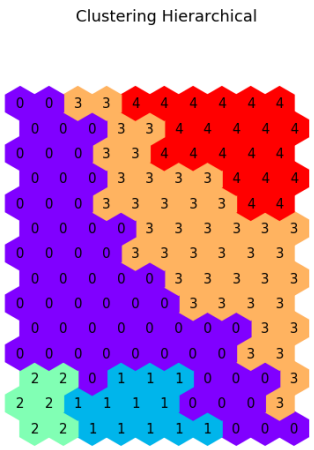

Figure 6.5: Barreiro Overview filtered data set - hierarchical clustering in SOM

From the two Figures (6.4 and 6.5), it is possible to see that both clustering techniques did almost an identical job at identifying the different clusters. This suggests that the different observations are correctly assigned to the respective group.

The boxplots of the different clusters to be analyzed are displayed below. Recall that the data was normalized to gather more meaningful insights regarding the values of each parameter in the clusters and to be able to represent them in the same Figures. Another view of these clusters displayed by parameter is illustrated in the Appendix Section.

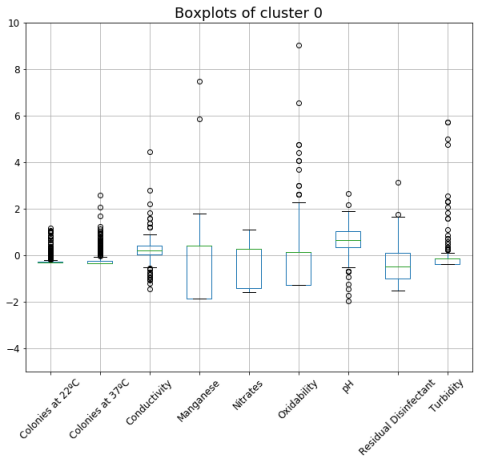

Figure 6.6: Barreiro Overview filtered data set - Cluster 0

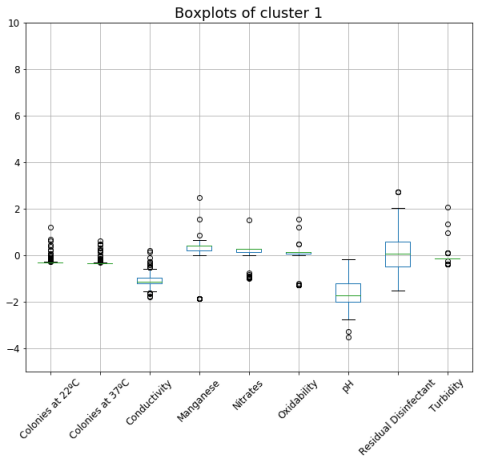

Figure 6.7: Barreiro Overview filtered data set - Cluster 1

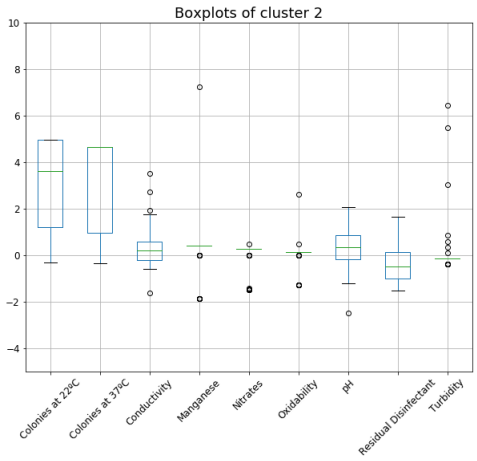

Figure 6.8: Barreiro Overview filtered data set - Cluster 2

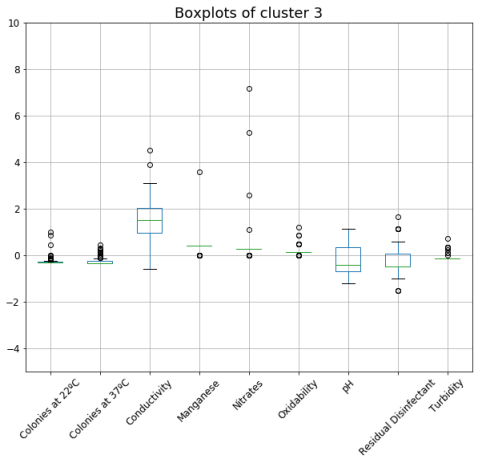

Figure 6.9: Barreiro Overview filtered data set - Cluster 3

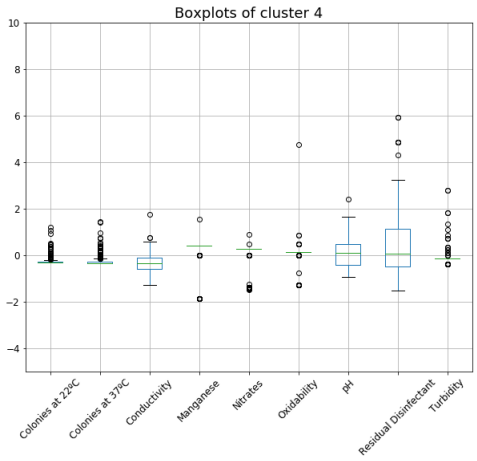

Figure 6.10: Barreiro Overview filtered data set - Cluster 4

Interpretation of the clusters obtained:

Starting with cluster 0, parameters such as the manganese, nitrates and oxidability attract the attention. For these three parameters, the variability of the data is bigger than in any other cluster, even if the median seems to be the same. This cluster also shows the biggest variation for the turbidity.

For cluster 1, the conductivity and pH shows little variation but its average values seem to be lower than the rest of the clusters.

Concerning cluster 2, the number of colonies at 22℃ and at 37℃ have the most variability and also the higher values of all the clusters, while the rest of the parameters seem to be quite average.

In regard to cluster 3, the conductivity has the most variability compared to all the other clusters as well as the highest average values.

Concerning cluster 4, the residual disinfectant is the most spread among all clusters, despite having the same median has cluster 1.

To conclude, observations with results off the average for the parameters manganese, nitrates and oxidability might be associated to cluster 0. Observations that show slim values for the conductivity and pH will tend to be associated to cluster 1. In the other way around, observations that demonstrate large values for either one of the number of colonies will have a much higher probability to be inserted into cluster 2. Looking at the conductivity, observations that show higher values of this parameter should be linked to cluster 3. Observations with values far from the average for the residual disinfectant might have a higher chance of being designated to cluster 4.

6.1.6 Water Quality Index

For the whole data set containing the parameters present in the document DL 152/2017 of the portuguese law, the three new proposed indexes were applied.

On the left, it is possible to see the different ratings and the values corresponding are the percentage of observations that fit in that category. Below the results, the mean of the WQI of the network is presented.

Regarding the first index, the results obtained are represented in Table 6.3.

Table 6.3: Rating classification for index A, B and C for the Barreiro data set

Rating	% of observations		
	Index A	Index B	Index C
Excellent	94.14	99.55	32.43
Very Good	-	-	1.35
Good	5.41	0.45	48.65
Poor	0.45	0	17.57
Very Poor	0	0	-
Unsuitable for drinking	0	0	0
Mean	11.37	4.47	29.93

Values marked with a ”-” mean that for that specific rating, the index does not have a classification defined. As it is observable, the main quantity of the observations are rated as excellent water, which is a very good sign for the water utility. One noticeable fact from this index is that there are no observations classified as unsuitable for drinking purposes nor rated as very poor. This highlights the quality of the water being delivered to the costumers for the period that there are analysis for this water utility. This fact contributes for such a good average of the network for this index. The average WQI for this index was 11.37. Overall, this is a very good sign since the water is classified as excellent.

Concerning the second index, the results obtained are actually quite similar to the first index. The percentage of observations for the different ratings can be seen in Table 6.3.In this case, the most noticeable changes from the classification in index A are the percentage of observations that are classified as excellent water that is higher and the inferior percentage of good water. There aren't any observations classified worse than very good, which is a very good sign. As expected, the average WQI for this index is significantly lower and has an overall result of 4.47.

For index C, the results obtained are also displayed in Table 6.3. For this index, it is much more difficult to actually make the difference between the good, very good and excellent quality of the water. The high amount of observations being classified as good derives from the fact that a lot of observations that are only tested for the 3 parameters inserted in the Routine Control 1 are within the parametric limits. This index allows a further exploration of why are there more observations classified as poor, which should allow the water utility to further investigate the situation/reasons why they are placed there, namely the parameter residual disinfectant not respecting its parametric values. There are still no observations classified as unfit for human consumption, which is in concordance with the other two indexes. The overall WQI for the Barreiro overview calculated using this index was 29.93. This labels the network as very good.

6.2 Beja

6.2.1 Statistical Analysis

Regarding the Beja data set, it contains all information regarding the parameter evaluations collected by EMAS Beja during the years 2009 until 2019. This data set has a total of 1807 observations and as many as 55 different parameters. The statistical characteristics are represented in Table 6.4. Here, in a similar fashion as it was done in the Barreiro data set, attributes such as the mean $\bar{x}$, standard deviation (s), minimum and maximum values, parametric value (PV) and number of observations for each parameter are displayed.

Table 6.4: Descriptive analysis of the parameters - Beja overall view

Parameter	$\bar{x}$	s	Min	Max	PV	no. Obs.
Colonies at 22℃*	9.82	31.66	0	301	100	614
Colonies at 37℃*	13.62	42.99	0	301	20	614
Conductivity	818.87	171.41	289	1700	2500	675
Hardness*	306.77	67.03	110	560	150-500	182
Iron	38.23	34.57	4.9	250	200	152
Manganese	14.56	8.74	0.49	160	50	667
Nitrates	18.21	13.59	1.9	67	50	205
Oxidability	1.46	0.61	0.49	4	5	667
pH	7.54	0.30	6.4	8.8	6.5-9.5	674
Residual Disinfectant*	0.29	0.21	0.02	2.1	0.2-0.6	1596
Temperature*	19.72	4.39	10.3	49.9	-	1093
THMs	20.57	29.23	0.1	147	100	152
TOC	3.44	1.16	1.3	6.7	WAA	30
Turbidity	0.52	0.57	0.19	5.6	4	673

Parameters noted with * are considered only recommended and not mandatory to comply with the legislation. WAA stands for without abnormal alteration. In a general way, most of the parameters are within their respective parametric values. The most studied parameter in this data set for these parameters only is the residual disinfectant. It also happens that this parameter has the most values being outside the respective recommended parametric interval. The number of colonies at 22℃ and at 37℃ are the parameters that show the largest dispersion of values with a significant portion of them being over the parametric value. In practical terms, the data set for these 14 parameters has a 1807 observations. The evolution of the parameters can be seen in the Appendix Section through boxplots.

6.2.2 Correlation Analysis

For the Beja data set, unlike the Barreiro one, all observations were utilized to perform the correlation matrix. This is due to the fact that, this time around there are 2 parameters with a lot of observations

(residual disinfectant and temperature), and like this it is possible to get a more accurate correlation result.

The correlation matrix is illustrated at Figure 6.11.

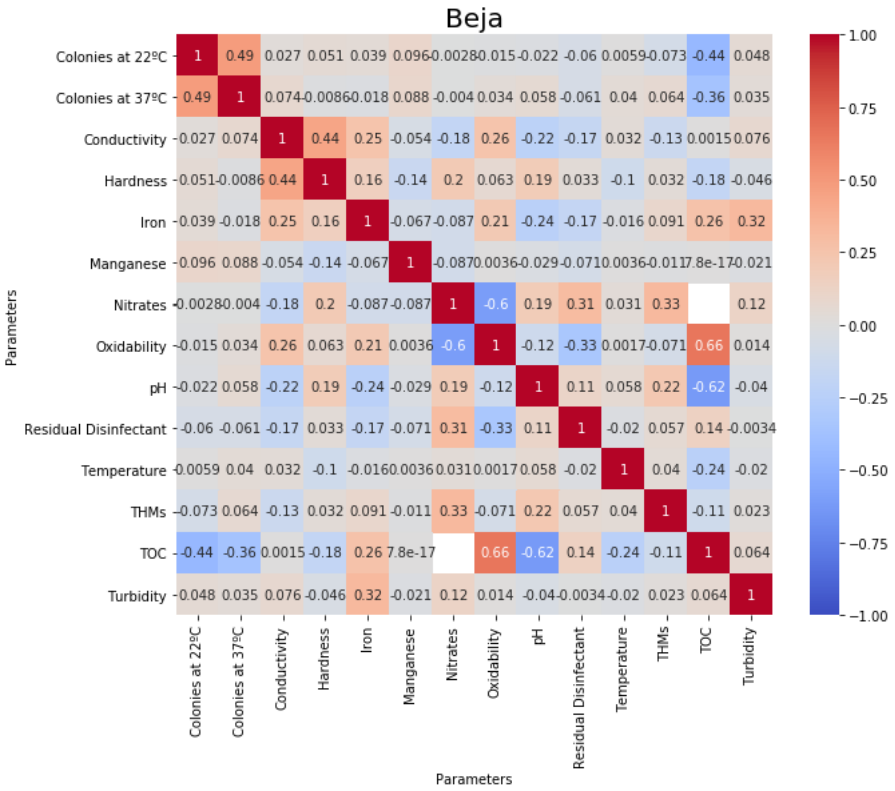

Figure 6.11: Correlation Matrix of the Beja Overview data set

From this graphic several conclusions can be drawn. Despite most of the correlations between the parameters are negligible, there 3 considered of importance. The pair oxidability and nitrates show a result of -0.6. This means that when one of them augments, the other one will show decreasing values. Another parameter that correlates highly with oxidability is TOC. This pair has a score of 0.66, and so correlates in a significant positive way. The last pair of parameters to show important value of -0.62 is the pair TOC and pH.

To further investigate these relations, a study by the years was produced. It can be seen in Figure 6.12.

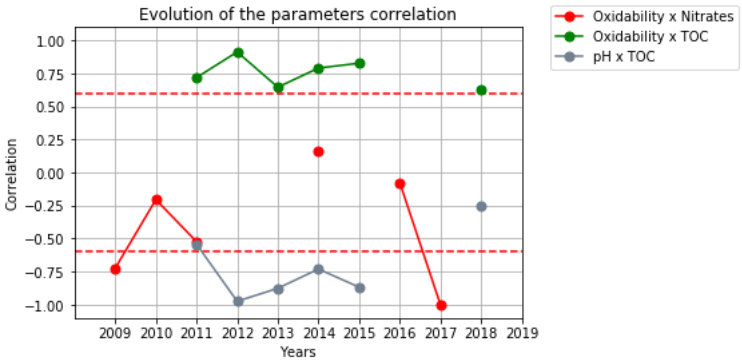

Figure 6.12: Correlation of parameters throughout time in Beja Overview data set

Observing how the relations evolve, it is possible to notice that all three pairs of parameters - oxidability/nitrates, oxidability/TOC and pH/TOC - behave differently. Starting with the first one, the oxidability and nitrates (red line), it is possible to see that the relation is very unstable throughout time is the value obtained (-0.6) is not a good reflection of how this pair actually behaves. The absence of values in the years 2012, 2013 and 2015 is a reflection of the constant value for the oxidability and so it was impossible to calculate the correlation for this pair those years.

For the next 2 correlations regarding TOC, they only have values in certain years because not all years contained information regarding this parameter.

Investigating the next case regarding the oxidability and TOC (green line), it is possible to conclude that it is much more stable and the overall value of this correlation is more accurate in what is shows. This suggests that increases in oxidability are most likely due to increases in TOC, either due to changes in upstream treatment or due to the intrusion of organic contaminants in the network.

As to the last one, the conclusions drawn can be similar, except for the year 2018, when the correlations hits a very small number. However, for the years when this relation was possible to calculate it showed consecutively high absolute values.

6.2.3 SOMs

For this analysis, it was necessary to trim down the data set. Because, only the temperature and the residual disinfectant have a very high number of observations, the data has a lot of missing values regarding the rest of the parameters. To perform the SOMs, there cannot be any missing values in the data set. Reminding that all parameters with over 66% of actual values will have its missing values replaced by their mean, while parameters with over 34% nonexistent values are to be removed, it was decided to only keep observations with two or more parameters studied. This came as a solution to discard the observations only containing 1 or 2 ”random” parameters and the observations only containing information concerning the residual disinfectant and temperature.

Regarding the percentage of missing values for each parameter, the hardness, iron, nitrates, THMs and TOC were removed from the filtered data set. All of the other parameters had some missing values and therefore had these values replaced by the their corresponding mean.

This is the filtered data set that will be used for the production of the SOMs. It includes 691 observations. So, the optimal number of nodes is near 131. The optimal map size chosen was a map of dimensions 13×10.

The best model trained has a topographic error of 0.0014 and a quantization error equal to 0.3858. The respective map is shown in Figure 6.13. The hits map of this model is displayed in the Appendix Section.

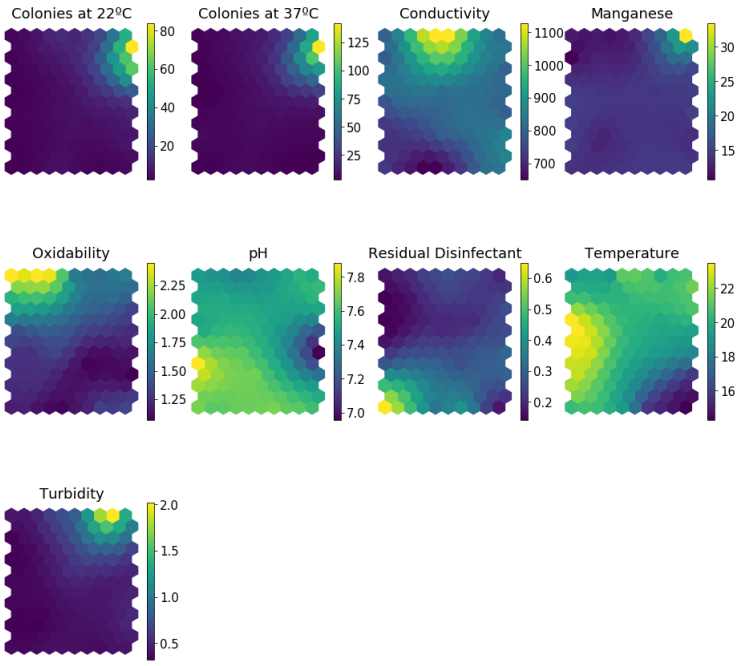

Figure 6.13: SOM of Beja Overview filtered data set

Observing the maps obtained, some relations are easily noticeable. Starting with the number of colonies at 22℃ and at 37℃, it is possible to perceive that they are both strongly correlated, meaning that high values of one parameter usually translates into also high values of the other. These two also parameters also correlate positively in weaker way with manganese and turbidity. Conductivity and oxidability also show a weak positive relation. The pH parameter correlates highly with temperature and slightly with residual disinfectant. This last one, is inversely correlated with both number of colonies, manganese and turbidity.

6.2.4 PCA

Regarding the PCA for the Beja data set, seven components were needed to accomplish 87.18% of the total variability of the original data set. The principal components explained a total variance of 19.01%, 16.55%, 11.55%, 11.31%, 10.31%, 9.85% and 8.61%, respectively. The Table 6.5 contains the correlations of the principal components with the selected parameters.

Table 6.5: Beja overview filtered data set - Principal components analysis

Parameter	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Colonies at 22℃	0.43	0.47	-0.13	-0.11	-0.24	-0.16	-0.15
Colonies at 37℃	0.46	0.48	0.07	-0.12	-0.16	-0.09	0.08
Conductivity	0.38	-0.37	-0.01	-0.34	0.03	-0.13	0.48
Manganese	0.15	0.15	-0.30	0.79	0.28	-0.15	0.37
Oxidability	0.35	-0.39	0.33	0.16	-0.27	0.49	0.32
pH	-0.24	0.36	0.36	0.11	-0.31	0.58	0.40
Residual Disinfectant	-0.37	0.26	-0.24	-0.41	0.18	-0.23	0.58
Temperature	0.10	0.19	0.69	-0.04	0.66	-0.15	-0.05
Turbidity	0.33	-0.01	-0.33	-0.18	0.44	0.72	-0.02

Concerning the first two principal components, it is possible to see they are quite similar. Both show a slight correlation with the number of colonies at 22℃ and at 37℃. The difference between them is the signal of the correlation with conductivity, pH and residual disinfectant. This leads to the conclusion that the first PC represents the relation of the number of colonies and the increase of conductivity, the alkaline levels and the decrease of residual disinfectant present in the water, while the second one represents the exact opposite relations with both number of colonies.

The third and fifth one is a estimate of the temperature since both have a high correlation with this parameter, suggesting this parameter is very important and has a big influence in the values of other parameters.

The fourth PC, correlates highly with manganese, showing that this component is a measure of the concentration of manganese of the samples.

The sixth parameter is a PC that correlates highly turbidity and pH. This is an indicator of the relation of these two parameters, showing that the rise of the values of one parameter result in an increase of the other.

The last and seventh PC correlates with residual disinfectant. This is a signal of the level of disinfectant present in the water.

6.2.5 Cluster Analysis

As for the cluster analysis, the optimal number of clusters was determined by an elbow plot and a dendrogram. Both revealed that the ideal value was three.

The representation of these clusters in the SOMs after the KMeans and hierarchical clustering was produced can be found in Figures 6.14 and 6.15.

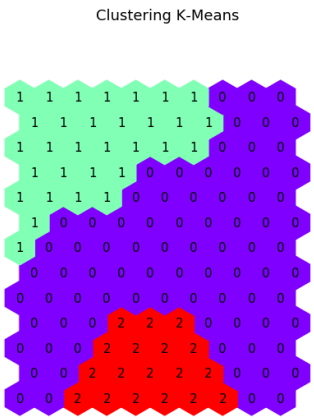

Figure 6.14: Beja Overview filtered data set - K-Means in SOM

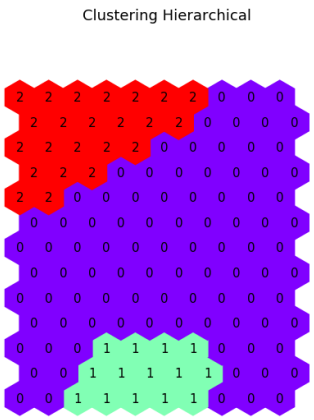

Figure 6.15: Beja Overview filtered data set - hierarchical clustering in SOM

It is noticeable that both clustering techniques did a very good job identifying correctly which observations placed in the maps belong to the correct clusters. This is concluded, since both maps are similar.

After normalizing the data for the parameters in question, the boxplots of each cluster were produced. These can be found in Figures 6.16, 6.17 and 6.18. A view of these clusters displayed by parameter can be found in the Appendix Section.

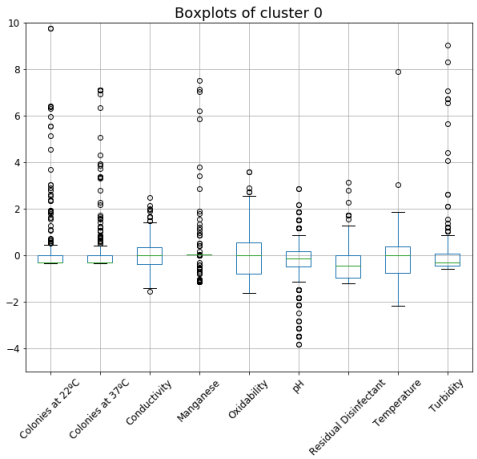

Figure 6.16: Beja Overview filtered data set - Cluster 0

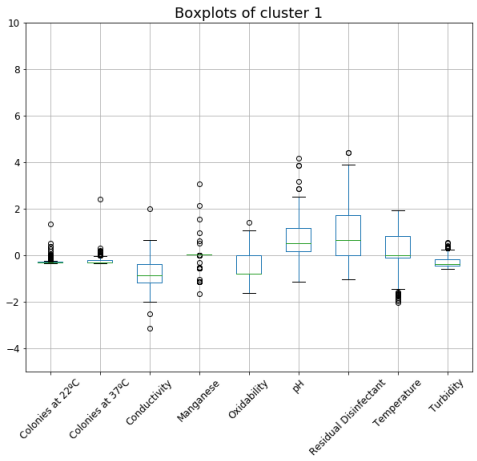

Figure 6.17: Beja Overview filtered data set - Cluster 1

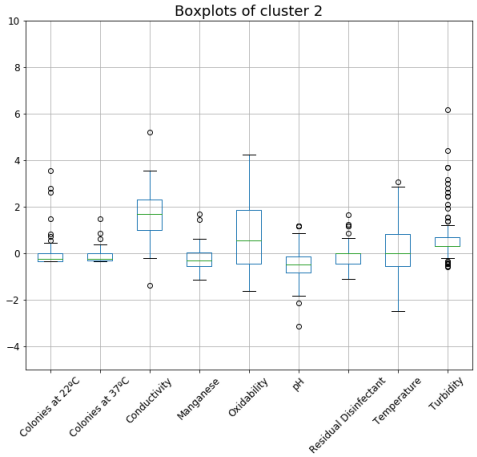

Figure 6.18: Beja Overview filtered data set - Cluster 2

Interpretation of the clusters obtained:

In the first cluster it is possible to find observations that show the largest variability with the number of colonies at both temperatures and manganese, temperature and turbidity. Observations in this cluster also have an average of lower values of residual disinfectant.

Regarding cluster 1, it shows usually lower levels of conductivity. However, the opposite occurs concerning pH and residual disinfectant showing higher values for these 2 parameters. Of all clusters, in this group it is noticeable that turbidity has the lowest variability.

As for the last cluster, cluster 2, it shows higher values and a big variability for the conductivity and oxidability.

Overall, it is possible to say that observations with higher values on the number of colonies, manganese, temperature and manganese tend to be located at cluster 0. The ones with large results for the pH and residual disinfectant and/or small values for the conductivity have a higher probability to be

associated to cluster 1. On the other hand, samples with great values of conductivity and oxidability should be found in cluster 2.

6.2.6 Water Quality Index

The new proposed indexes were calculated for this data set. Once again, the percentage of observations is shown for each rating and the last line is designated for the mean of the index calculated.

For the index A, the results obtained can be found in Table 6.6.

Table 6.6: Rating classification for index A, B and C for the Beja data set

Rating	**% of observations**		
	Index A	Index B	Index C
Excellent	94.63	98.12	26.45
Very Good	-	-	10.63
Good	3.65	0.22	23.08
Poor	0.06	0	38.20
Very Poor	0	0	-
Unsuitable for drinking	1.66	1.66	1.61
Mean	15.36	12.09	29.93

Recall that values marked with a ”-” mean that for that specific rating, the index does not have a classification defined. The water in this water utility is classified overall as excellent with a mean of 15.3620. The vast majority of the observations are ranked as excellent and only a very small percentage is classified as not suitable for human consumption. This is mainly due to the flagged observations containing any harmful microbiological activity.

Concerning the second index proposed, or index B, the results are displayed in Table 6.6 The results obtained for index B are quite similar to the ones acquired for index A. Result of having more observations classified as excellent, results in a decrease of the mean of this WQI. It has the same percentage of water unsuitable for drinking as index A, suggesting that the observations placed in this rating are the same for both indexes, with the same reasons.

As for the index C, the scores are in Table 6.6. The classifications obtained for this index rate the water between good and very good. This is to be expected, since the water shows good quality wit its parameters within parametric values most of the time. There is a large percentage of observations evaluated as poor, due to the fact of residual disinfectant not respecting its parametric value. Despite this fact, the mean of the WQI for this water network is still pretty good and almost rated as very good.

6.3 Infraquinta

6.3.1 Statistical Analysis

Finally, the last data set mentioned in this work is the the Infraquinta data set collected by the company in Quinta do Lago, Portugal from the years 2008 up to 2019 with the exception of 2011. Once again, the original data set was transformed in a way to make each observation a pair sampling site and date of analysis and each parameters was considered a variable. This data set contained only 262 observations and a total of 35 distinct parameters. The principal statistical analysis of this data for the principal parameters collected by Infraquinta is displayed in Table 6.7.

Table 6.7: Descriptive analysis of the parameters - Infraquinta overall view

Parameter	$\bar{x}$	s	Min	Max	PV	no. Obs.
Colonies at 22℃*	14.65	57.43	0	301	100	200
Colonies at 37℃*	11.49	52.22	0	301	20	200
Conductivity	235.98	69.60	0	750	2500	200
Hardness*	93.64	25.98	2.7	150	150-500	32
Iron	23.83	23.44	0.01	90	200	32
Manganese	7.05	10.82	0.49	123	50	123
Oxidability	1.14	0.40	0.49	4.5	5	201
pH	7.71	0.28	6.7	8.8	6.5-9.5	200
Residual Disinfectant*	0.37	0.19	0.09	1	0.2-0.6	261
THMs	37.60	13.91	17.83	69.63	100	32
Turbidity	0.50	0.48	0.39	7	4	201

Recall that parameters noted with * are considered only recommended and not mandatory to comply with the legislation. Most of the observations seem to have all these selected parameters within parametric values. The residual disinfectant shows a large number of observations over the parametric value. This parameter is also the one tested more often, as it was the case in the previous 2 case studies. This is due to this parameter belonging in the group Routine Control 1. The three groups are easily identifiable by the number of observations of each parameter. Another interesting result is the distribution of the hardness values. Almost all of the observations that contain information concerning this parameter are below the minimum recommended value. This could be explained by the natural properties of the water in that region and by the fact that this parameter is only recommended and not mandatory. There are 262 observations for the 11 chosen parameters. The statistical evolution of these parameters are displayed in the Appendix Section.

6.3.2 Correlation Analysis

The data set used to perform the statistical analysis was the one utilized to perform the correlation study. The results obtained are displayed in the Figure 6.11.

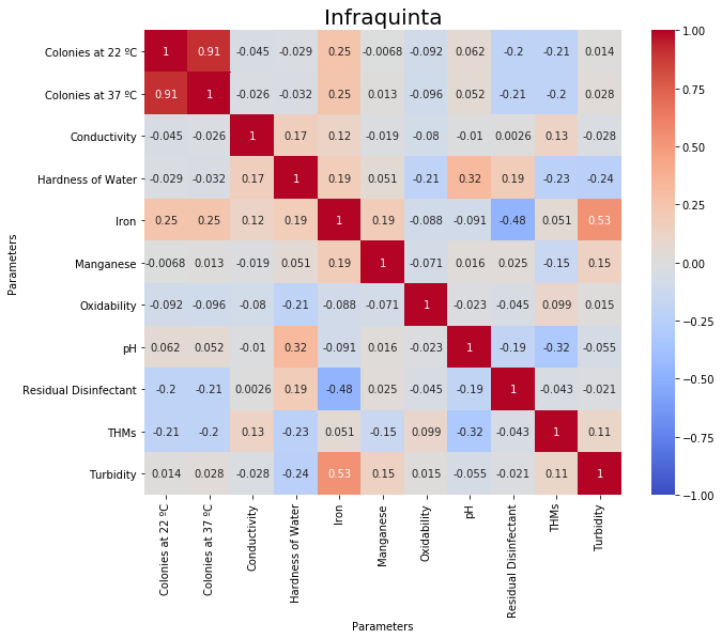

Figure 6.19: Correlation Matrix of the Infraquinta Overview data set

Looking at results obtained there is only one of important. The high value for the relation between the number of colonies at 22 ºC and at 37 ºC shows that the high score of one parameter is highly correlated to large values of the other. This suggests that microorganisms present in this water distribution network are able to grow at higher temperatures of 37 ºC and so, able to infect people. The evolution of this relation can be further seen in Figure 6.20.

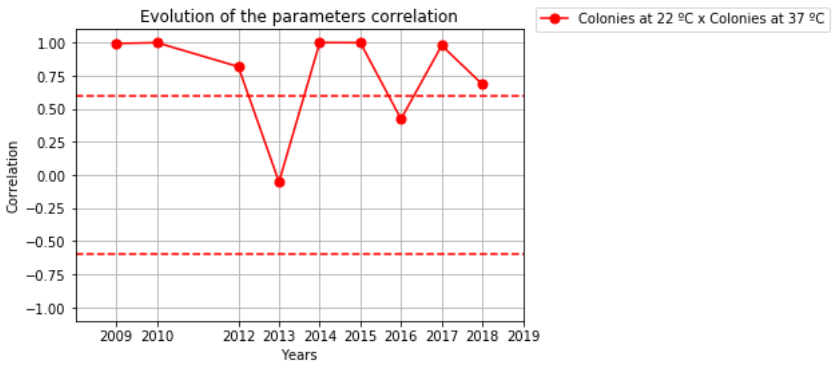

Figure 6.20: Correlation of parameters throughout time in Infraquinta Overview data set

For all years that there are observations in this data set, it is clear that only in the year of 2013 the correlation of both parameters was close to null. Apart from this year and 2016, the correlation values were always above around 0.7 and often had a result of close to 1. Overall, this relations seems quite steady and so the value obtained for the heatmap (Figure 6.19) and the conclusions drawn are justified.

6.3.3 SOMs

Once again, to perform the SOMs, some changes had to be done to the data set. Here, since the residual disinfectant contained some more values than all the other parameters, it was decided to remove all observations that only tested this 1 parameter. This was done for two reasons. The first one to be consistent with the number of missing values for all the other parameters and the second one was just because this would not affect the results for the maps, since it was impossible to calculate correlations between the residual disinfectant and any other parameter for such observations.

Since parameters with a percentage of missing values over 34% are removed, the filtered data set of Infraquinta has 3 less parameter: hardndess, iron and THMs. For the other parameters, their missing values are replaced by their respective mean. This filtered data set has 201 observations and 8 parameters.

For this amount of observations, the optimal number of nodes for the map is near 70 nodes. This results in a map with 10×7 dimensions. The best model trained has a topographic error of 0.0 and a quantization one of 0.3064. This map is shown in Figure 6.21. The corresponding hits map can be found in the Appendix Section.

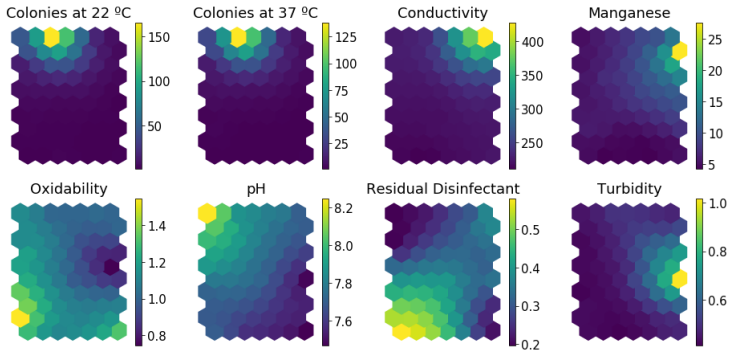

Figure 6.21: SOM of Infraquinta Overview filtered data set

From the SOM obtained for this data set, further conclusions and relations between the parameters can be seen. Confirming what was in the correlation heatmap the relation between both number of colonies at 22 ℃ and at 37 ℃ is noticeable. They both show high values for the same regions of the map. They both show a slight positive relation with pH and an inverse correlation with residual disinfectant and turbidity. Conductivity and manganese also show a little positive correlation. In addiction to these relations, oxidability and residual disinfectant have a somewhat strong positive relation.

6.3.4 PCA

For this data set, to explain 89.58% of the total variance, it was necessary to have 6 components. They, respectively, account for 25.36%, 14.76%, 13.92%, 13.49%, 11.84%, 10.22% of this variance. The correlation of these components with the original parameters are found at Table 6.8.

Table 6.8: Infraquinta overview filtered data set - Principal components analysis

Parameter	PC1	PC2	PC3	PC4	PC5	PC6
Colonies at 22℃	0.67	0.06	-0.11	-0.14	-0.01	0.08
Colonies at 37℃	0.67	0.08	-0.11	-0.13	0.02	0.09
Conductivity	-0.04	-0.02	-0.45	0.47	0.71	0.24
Manganese	0.01	0.57	0.31	0.35	-0.19	0.62
Oxidability	-0.11	-0.28	0.45	-0.50	0.38	0.51
pH	0.13	-0.40	0.35	0.56	-0.25	-0.03
Residual Disinfectant	-0.29	0.35	-0.40	-0.24	-0.28	0.15
Turbidity	0.03	0.54	0.44	-0.01	0.42	-0.53

The first PC correlates highly with the number of colonies at 22℃ and at 37℃. This suggests that this component is a measure of the microorganisms present in the water.

The second one has a high relation with the manganese and turbidity and could be seen as a representation of the presence of these two parameters in the water.

The third PC does not have a very high relation with any parameter. The two most significant ones are the values obtained for the conductivity (-0.45) and oxidability (0.45). Since the values are not very representative it is difficult to draw any conclusions on what this PC is really representing.

The next PC correlates highly with pH and so it can be a sign of how alkaline the water is.

Principal component 5 scored the highest with conductivity. This leads to the believe that this PC is an indication of how conductive the water is in the observations.

The sixth and final component correlates highly with two parameters (manganese and oxidability) and quite low with turbidity. So, this PC can be seen as a measure of these 3 parameters. While a high positive value for the component leads to high values of the manganese and oxidability, for the turbidity it leads to low values.

6.3.5 Cluster Analysis

Regarding the cluster analysis, the elbow plot from K-Means and the dendrogram from the hierarchical clustering techniques were put into place. Both methods reached the same amount of optimal clusters of 3. The position of this clusters in the SOMs for each one of the techniques is shown in Figures 6.22 and 6.23.

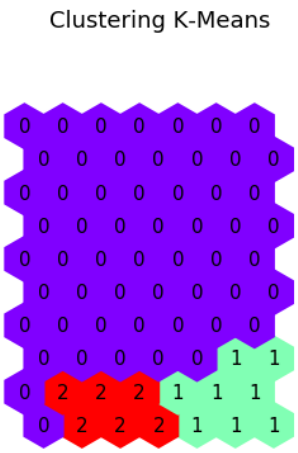

Figure 6.22: Infraquinta Overview filtered data set - K-Means in SOM

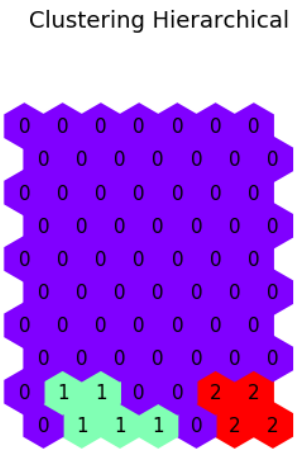

Figure 6.23: Infraquinta Overview filtered data set - hierarchical clustering in SOM

In general, the position of the clusters identified by both methods was similar, but some differences are quite noticeable. It is notorious that cluster 0 occupies the most hexes in both methods. However, in the K-Means technique cluster 1 and 2 have a larger importance than in the hierarchical clustering.

Using the clusters created by the K-Means technique, the boxplots of each cluster are displayed in Figures 6.24, 6.25 and 6.26. Another perspective where the parameters are displayed by clusters can be seen in the Appendix Section.

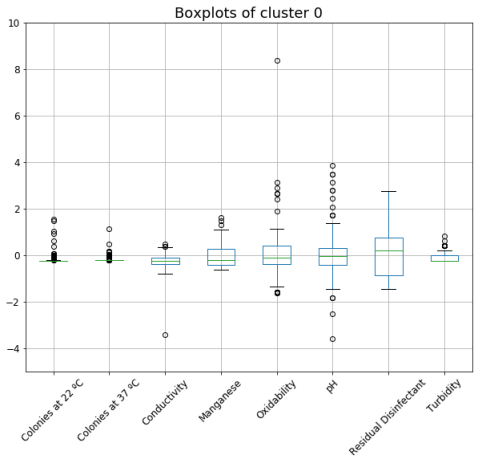

Figure 6.24: Infraquinta Overview filtered data set - Cluster 0

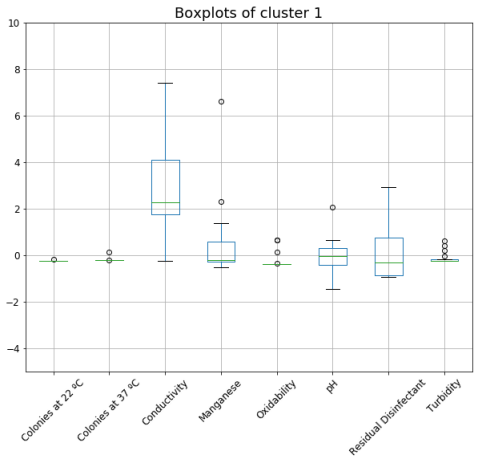

Figure 6.25: Infraquinta Overview filtered data set - Cluster 1

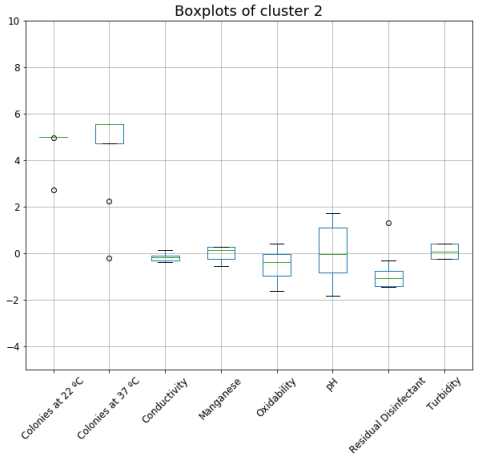

Figure 6.26: Infraquinta Overview filtered data set - Cluster 2

Interpretation of the clusters obtained:

Regarding cluster 0, the conductivity shows a slightly wide distribution but with overall lower values than the other clusters. The oxidability, pH and turbidity show the widest distribution of data of all clusters.

As for cluster 1, it has the largest variability and the higher values of conductivity in the data of all clusters. This cluster also shows the greatest variability for the manganese. This cluster also has the smallest variability of data for the oxidability.

Regarding the final cluster, it has the highest medians for both number of colonies at 22℃ and at 37℃ and show the largest variability for this last one. It also shows lower levels and the thinnest variability for the residual disinfectant.

To conclude, in cluster 0 it is possible to find observations containing low values for the conductivity and a wide range of results of oxidability, pH and turbidity. In cluster 1, it is likely to find high values for

the conductivity and manganese. Lastly, high values for the number of colonies are associated to cluster 2.

6.3.6 Water Quality Index

In this section, the results obtained for the 3 proposed WQI for the Infraquinta data set are shown. Once again, the rating are displayed first and in the last row of the Table the mean of the index is represented. These results are found in Table 6.9.

Table 6.9: Rating classification for index A, B and C for the Infraquinta data set

Rating	% of observations		
	Index A	Index B	Index C
Excellent	98.09	97.71	58.02
Very Good	-	-	18.32
Good	0.76	1.91	13.36
Poor	0.38	0	9.54
Very Poor	0	0	-
Unsuitable for drinking	0.76	0.38	0.76
Mean	22.35	12.48	41.84

Keep in mind that values marked with a ”-” mean that for that specific rating, the index does not have a classification defined. Observing the obtained results for the index A, it is possible to see that the results are very good as the quality of the water for this index is classified as excellent. Almost all observations are rated as excellent, proving that the water quality control that as been done for this water utility is impressive. There are, however, two observations where the water is considered unsuitable for drinking purposes. The reasons behind this is one flagged observation that contained microbiological activity and the other had high values for several parameters, namely the conductivity that showed a significant high value.

Concerning the second index proposed, or index B, the results are displayed in Table 6.9. For this index, similar results as index A were found. The water is still classified as excellent, with an even better overall mean than the one found in index A. The main difference is due to the observation with the odd conductivity level, that was, for this index, classified as excellent. This illustrates the differences between WQI and how they can impact the view of the entire network. The observation with water correctly identified as unsuitable for human consumption is still the flagged one, already mentioned in index A.

As for the index C, the scores are in Table 6.9. Looking at the percentages of each rating and the mean obtained, the results are consistent with the other indexes A and B. The mean obtained classifies this water network with a rating of excellent, which is a good sign for the water utility. There is a larger spread of observations classified as poor and good, unlike the other 2 new indexes proposed. The ”poor”

water has a significant percentage and represents mainly the parameters from Routine Control 1 (E.coli, coliform bacteria and residual disinfectant), where the residual disinfectant was not within the parametric recommended values. On the other hand, the ”good” water is represented by the observations where the Routine control 1 was done, but had all parameters within parametric values. The two observations ranked with water improper for human consumption are the flagged one with microbiological activity and another one from Routine Control 1, where the residual disinfectant was very off the recommended value and so should be unfit for people to drink.

7

Conclusion

Contents

7.1 Conclusions

In this section, a recap of the conclusions drawn from the analysis is accomplished. This is a review of all the outcome of the techniques applied for each network summed up. For each type of analysis, a summary of all results obtained is presented, along with the common and different things between the case studies. Lastly, a section referring to future work is presented.

Correlation Analysis:

The most notorious correlations found in Barreiro after a more detailed study were the relation between the number of colonies at 22℃ and at 37℃ and the relation between hardness and conductivity. These tend to show a steady correlation above the 0.6 threshold across almost every year.

As for the Beja data set, three other relations were investigated due to their high correlation. Here, the oxidability and nitrates showed an negative correlation. However, upon a study across the time there were analysis, this relation was neglected. Oxidability and TOC showed a high correlation steady along time, as well as the relation between pH and TOC. On the other hand, this last one displayed a negative correlation value.

The final data set, gathered by Infraquinta, showed only one very high correlation regarding the number of colonies at 22℃ and at 37℃. After a further examination, it was concluded that this was a very high value across almost every year.

The relation between the number of colonies at 22℃ and at 37℃ was to be expected, since it makes sense from the microbiological sense. If there are microorganisms at 22℃, it becomes clear that an augment on this parameter would influence positively the presence of harmful microorganisms at 37℃. This is easily seen in the Barreiro and Infraquinta data set. Some other relations such as the one between hardness and conductivity as seen in the Beja data set, are logical from the physical point of view, since hardness corresponds to the sum of calcium and magnesium ions and so the conductivity indicates the capacity of the water to conduct electricity due to those ions.

SOM analysis:

Concerning the analysis of the SOMs, some further conclusions were drawn. Starting with the Barreiro data set, it was clear that there was a positive relation with both number of colonies. These ones were inversely related to residual disinfectant, which makes sense from the physical point of view. Some other weaker relations were found, namely the relation between the parameters nitrates, oxidability, manganese and conductivity.

Regarding the Beja data set, the main relations are found between the number of colonies at 22℃ and at 37℃. Once more, these parameters have an inverse relation with residual disinfectant. They also show a feeble relation with manganese and turbidity. It is also possible to observe positive correlations between pH, temperature and residual disinfectant.

In the Infraquinta data set, once again, both number of colonies at 22℃ and at 37℃ relate with each

other and both are inversely associated to the residual disinfectant and turbidity. The parameter of the colonies also show an association with pH. There are 2 other links of importance. The first one being the association of the conductivity and manganese. The second one is the positive correlation between oxidability and residual disinfectant.

Taking an overall look at 3 conclusions, it is perceivable in all 3 locations that the number of colonies at 22℃ and at 37℃ are correlated, and they both appear to be inversely associated to residual disinfectant. In Beja and Infraquinta, the number of colonies appear to be also inversely linked to turbidity. The other associations tend to be related to specifics of each data set and how the data collected from each location behaves.

Principal Component Analysis (PCA):

Overall, to have around 85%-90% of the variability of the data set explained by principal components, each case study has 6 or 7 components. This number is to be expected since there are usually 8 or 9 parameters being studied when this kind of analysis is performed and each PC correlates normally with 1 or 2 parameters.

The principal component common in all the case studies is one component that correlates highly with the number of colonies at 22℃ and at 37℃. One justification for this is the fact that these two parameters already correlate highly with each other, as seen in the previous analysis of correlation and SOMs.

Cluster Analysis:

Regarding the cluster obtained after performing the SOMs, some interesting results were found. The Barreiro data set was the only one that had an optimal number of clusters of 5, instead of the 3 optimal for the other case studies. This is mainly due to the superior number of observations in this case and the consequent increase of complexity in the data.

The most noticeable common cluster is one carrying high variability and high average and median levels for the number of colonies at 22℃ and at 37℃ that was present in all cases. Conductivity was also a major key factor in the building of the clusters. This parameter was a determinant factor for the description of one or more clusters as for these cases it showed a higher or lower variability and/or usually higher or lower values. In other words, there were clusters built around this parameter that was the only difference compared to the rest of the data. In the Barreiro data set, cluster 3 showed a high variability and usually higher levels for this parameter. The same occurred for Beja in cluster 2 and for Infraquinta in cluster 1. This parameter was also involved in the description of cluster 1 in the Beja case study as it showed little variability. Another parameter that was often involved in the and seen as a key factor in the clusters was the residual disinfectant. In the Barreiro data set, cluster 4 showed high variability for this parameter, while in the Infraquinta case study cluster 2 displayed little variability, while associated to high levels of number of colonies at both temperatures. For the Beja data set, this

parameter appears in 2 clusters. The first time, in cluster 0, showing lower than usual values, while associated to large a variability of the number of colonies, and at cluster 1, where it showed larger values. These larger values were associated to great values of pH, lower average values of conductivity and a small variability for turbidity.

WQI

Concerning the water quality indexes, an evaluation of the three new proposed indexes for each one of the case studies is presented.

Regarding index A, the lowest mean was detected for Barreiro (11.37), followed by Beja (15.36) and then Infraquinta (22.35). This suggests that the best quality of the water is found in these places in that order. Even though there are some differences between the data sets, all 3 were classified as having excellent water quality. As for index B, that has the same rating scale of index A, the order of the means calculated for this index remain the same, but now on a different scale. Overall, Barreiro still shows the best water quality with an average of 4.47, followed by Beja (12.09) and then Infraquinta (12.48). The water is still classified as excellent for the 3 locations for this index and the values obtained are significantly lower than the ones obtained for index A. This suggests that this index benefits more the quality of the water than index A. With regard to the last index, index C, the results were slightly different from the ones obtained for the past two WQIs, namely because of the inclusion of the residual disinfectant and the different classification method applied. For this case, the case study considered of having the best water quality is Infraquinta with an overall score of 41.84. The second best water quality is found in the Barreiro data set where the mean was 29.93. These two water utilities have a rating of very good water, according to the ratings defined. Lastly, Beja had a result of 24.31. This score is on the verge of being considered very good, but nevertheless it is rated as good.

The WQIs calculated for the three data sets showed outstanding results. It is remarkable that there was not detected any microbiological activity (E.Coli, coliform bacteria or enterococci) in the Barreiro water utility, resulting in 0 flagged observations. Thus, the very low results for the first two indexes. The Beja data set showed very similar results to Barreiro, but had a few observations that diminuished the score obtained, either by having flagged observations or by having several parameters outside the parametric values. As to the Infraquinta data set, despite having a poorer results for the first two indexes, index C displayed very interesting results. By having a lot more observations ranked as excellent, the mean of this index for this location presented a massive increase. This means that for this water utility, there were made a lot more observations where many parameters were evaluated, all within parametric values. It also shows the lowest percentage of poor quality water, meaning that the residual disinfectant is more controlled and under recommended values than in all the other two cases being studied. This could be a sign that Infraquinta pays more attention to the recommended parameters than the other water utilities and is more focused on keeping them under or within their respective recommended

values.

To conclude this work, the three water utilities revealed to have some things in common, as seen in the correlation and SOMs analysis, but it was clear that each data set/location has its own particularities and so it was important to make individual reports about each one of them as they presented very different and curious results. All three of them also displayed to have very good or excellent water quality in their network, confirming that the water being delivered to the customers is of very high quality and that the control and management of the network is of an impressive standard.

It could have been useful to have more data available across more years of analysis as well as more parameters being evaluated more often, despite the minimum number described in the legislation. The inclusion of parameters such as temperature, as was the case in Beja, could have been of great interest as it is a very easy to measure, not costly and of great value to see how it impacts all the other parameters. Nevertheless, it was interesting to see how the use of different indexes changes the overall water quality being presented to the costumers.

7.2 Future Work

There is clearly more to be done in this area of analysis of water quality, where new techniques and new optics would be of a massive benefit. Applying new techniques and more advanced techniques besides the SOMs could be a great addition since new information and relations could be found and analyzed. This new methods could be very interesting to apply, specially if they are related to ANNs that have been showing a very good performance in the last years. Regarding the relations between parameters, it could be curious to see if it was possible to deduce the value with certitude of one or more parameters from others across the network. This would allow the water utilities to save money by not having to analyze everything every time, while still having a clear idea of the values of the parameters in the network.

It would also be of importance to see how the clusters are displayed in the water network and see which regions have a similar water quality and investigate the reasons behind the clustering processes.

Along the same train of thought, having the water quality value calculated by WQIs displayed in the network could bring new insights of why some sampling sites have a good or bad index score and how they relate to the rest of the network. It could bring another perception of why some places or regions have the same results. It could also be possible to detect any choke points than influence the rest of the network that goes beyond that point and to detect if there are any trends of sampling sites having parameters not respecting their respective parametric values after some time.

Bibliography

[1] A. Dertat, "Applied deep learning - part 1: Artificial neural networks," https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6, May 2021.

[2] "The ultimate guide to self organizing maps (som's)," https://www.superdatascience.com/blogs/the-ultimate-guide-to-self-organizing-maps-soms, May 2021.

[3] C. Thomas, "Water in crisis: a guide to the world's fresh water resources," *International Affairs*, vol. 70, no. 3, pp. 557–557, 07 1994. [Online]. Available: https://doi.org/10.2307/2623756

[4] F. Caroline, "he Impact of Climate Change: The World's Greatest Challenge in the Twenty-first Century, New Holland Publishers Ltd ," 2008.

[5] "Wisdom project," https://wisdom.ips.pt/, November 2020.

[6] S. Palani, S.-Y. Liong, and P. Tkalich, "An ann application for water quality forecasting," *Marine pollution bulletin*, vol. 56, pp. 1586–97, 09 2008.

[7] P. Juntunen, M. Liukkonen, M. Lehtola, and H. Yrjö, "Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process," *Applied Soft Computing*, vol. 13, p. 3191–3196, 07 2013.

[8] Y. An, Z. Zou, and R. Li, "Descriptive characteristics of surface water quality in hong kong by a self-organising map," *International Journal of Environmental Research and Public Health*, vol. 13, p. 115, 01 2016.

[9] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, pp. 1464–1480, 10 1990.

[10] W. Wu, Kui Chang, J. Gao, Min Zhang, Nana Li, and Yixing Yuan, "Research on water quality comprehensive evaluation of water supply network using som," pp. 714–718, 2009.

[11] W. Wu, G. Dandy, and H. Maier, "Protocol for developing ann models and its application to the assessment of the quality of the ann model development process in drinking water quality modelling," *Environmental Modelling Software*, vol. 54, p. 108–127, 04 2014.

[12] A. Kalteh, P. Hjorth, and R. Berndtsson, "Review of the self-organizing map (som) approach in water resources: Analysis, modelling and application," *Environmental Modelling Software*, vol. 23, no. 7, pp. 835–845, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815207001879

[13] S. Mounce, I. Douterelo, R. Sharpe, and J. Boxall, "A bio-hydroinformatics application of self-organizing map neural networks for assessing microbial and physico-chemical water quality in distribution systems," 07 2012.

[14] S. Mounce, R. Sharpe, V. Speight, B. Holden, and J. Boxall, "Knowledge discovery from large disparate corporate databases using self-organising maps to help ensure supply of high quality potable water," 08 2014.

[15] S. Mounce, E. Blokker, S. Husband, W. Furnass, P. Schaap, and J. Boxall, "Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems," *Journal of Hydroinformatics*, vol. 1, 01 2016.

[16] E. Blokker, W. Furnass, J. Machell, S. Mounce, P. Schaap, and J. Boxall, "Relating water quality and age in drinking water distribution systems using self-organising maps," *Environments*, vol. 3, p. 10, 04 2016.

[17] V. Speight, S. Mounce, and J. Boxall, "Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets," *Environmental Science: Water Research Technology*, vol. 5, 01 2019.

[18] G. Kyritsakas, V. Speight, S. Mounce, and J. Boxall, "Investigating drinking water behaviour treated by different disinfection with the use of a machine learning technique on water quality datasets," *17th International Computing Control for the Water Industry Conference*, 10 2019.

[19] J. Yisa and J. Tijani Oladejo, "Analytical studies on water quality index of river landzu," *American Journal of Applied Sciences*, vol. 7, 04 2010.

[20] M. Rahman, T. Akter, F. Jhohura, F. Akter, T. Chowdhury, S. K. Mistry, D. Dey, M. Barua, and M. Islam, "Water quality index for measuring drinking water quality in rural bangladesh: A cross-sectional study," *Journal of Health Population and Nutrition*, vol. 201635:4, 02 2016.

[21] S. Tyagi, P. Singh, B. Sharma, and R. Singh, "Assessment of water quality for drinking purpose in district pauri of uttarakhand, india," *Applied Ecology and Environmental Sciences*, vol. 2, pp. 94–99, 01 2014.

[22] A. Al-Afify, A. Othman, and M. Hassanien, “Characterization of chemical and microbiological quality of nile river surface water at cairo (egypt),” *Rendiconti Lincei. Scienze Fisiche e Naturali*, vol. 29, 06 2018.

[23] S. Bouslah, L. Djemili, and L. Houichi, “Water quality index assessment of koudiat medouar reservoir, northeast algeria using weighted arithmetic index method,” *Journal of Water and Land Development*, vol. 35, 12 2017.

[24] S. Dutta, A. Dwivedi, and M. S. Kumar, “Use of water quality index and multivariate statistical techniques for the assessment of spatial variations in water quality of a small river,” *Environmental Monitoring and Assessment*, vol. 190, 11 2018.

[25] M. Ibrahim, “Assessing groundwater quality for drinking purpose in jordan: Application of water quality index,” *Journal of Ecological Engineering*, vol. 20, pp. 101–111, 03 2019.

[26] A. Badeenezhad, H. Tabatabaee, H.-A. Nikbakht, M. Radfard, A. Abbasnia, M. Baghapour, and M. Alhamd, “Estimation of the groundwater quality index and investigation of the affecting factors their changes in shiraz drinking groundwater, iran,” *Groundwater for Sustainable Development*, vol. 11, p. 100435, 06 2020.

[27] S. Olasoji, N. Oyewole, B. Abiola, and J. Edokpayi, “Water quality assessment of surface and groundwater sources using a water quality index method: A case study of a peri-urban town in southwest, nigeria,” *Environments*, vol. 6, p. 23, 02 2019.

[28] A. Alomran, F. Al-Barakah, A. Altququ, A. Aly, and M. Nadim, “Drinking water quality assessment and water quality index of riyadh, saudi arabia,” *Water Quality Research Journal of Canada*, vol. 50, pp. 287–296, 08 2015.

[29] A. Alver, “Evaluation of conventional drinking water treatment plant efficiency according to water quality index and health risk assessment,” *Environmental Science and Pollution Research*, 09 2019.

[30] C. Almeida, S. González, M. Mallea, and P. González, “A recreational water quality index using chemical, physical and microbiological parameters,” *Environmental science and pollution research international*, vol. 19, pp. 3400–11, 04 2012.

[31] H. Boyacioglu, “Development of a water quality index based on a european classification scheme,” vol. 33, 02 2007.

[32] C. Cude, “The oregon water quality index (owqi) - a communicator of water quality information,” 04 2001, pp. 125–137.

[33] H. Khan, A. Khan, and S. Hall, “The canadian water quality index: a tool for water resources management,” 01 2005.

[34] T. Hurley, R. Sadiq, and A. Mazumder, “Adaptation and evaluation of the canadian council of ministers of the environment water quality index (ccme wqi) for use as an effective tool to characterize drinking source water quality,” *Water research*, vol. 46, pp. 3544–52, 04 2012.

[35] M. Mohebbi, R. Saeedi, A. Montazeri, K. Vaghefi, S. Labbafi, S. Oktaie, M. Abtahi, and A. Mohagheghian, “Assessment of water quality in groundwater resources of iran using a modified drinking water quality index (dwqi),” *Ecological Indicators*, vol. 30, p. 28–34, 07 2013.

[36] K. L. de Oliveira, R. L. Ramos, S. C. Oliveira, and C. Christofaro, “Water quality index and spatio-temporal perspective of a large Brazilian water reservoir,” *Water Supply*, vol. 21, no. 3, pp. 971–982, 12 2020. [Online]. Available: https://doi.org/10.2166/ws.2020.374

[37] A. R. M. Islam, A. Al Mamun, M. M. Rahman, A. Zahid, and R. Dufault, “Simultaneous comparison of modified-integrated water quality and entropy weighted indices: Implication for safe drinking water in the coastal region of bangladesh,” *Ecological Indicators*, vol. 113, 03 2020.

[38] V. Amiri, M. Rezaei, and N. Sohrabi, “Groundwater quality assessment using entropy weighted water quality index (ewqi) in lenjanat, iran,” *Environmental earth sciences*, vol. 72, pp. 3479–3490, 04 2014.

[39] H. Gharibi, M. H. Sowlat, A. Mahvi, H. Mahmoudzadeh, H. Arabalibeik, M. Keshavarz, N. Karimzadeh, and G. Hassani, “Development of a dairy cattle drinking water quality index (dcwqi) based on fuzzy inference systems,” *Ecological Indicators*, vol. 20, p. 228–237, 09 2012.

[40] A. Sutadian, N. Muttil, A. Yilmaz, and B. Perera, “Using the analytic hierarchy process to identify parameter weights for developing a water quality index,” *Ecological Indicators*, vol. 75, pp. 220–233, 04 2017.

[41] S. Mukhopadhyay, S. Maiti, and R. Masto, “Development of mine soil quality index (msqi) for evaluation of reclamation success: A chronosequence study,” *Ecological Engineering*, vol. 71, p. 10–20, 10 2014.

[42] R. Patterson, P. Haines, and B. Popkin, “Diet quality index: Capturing a multidimensional behavior,” *Journal of the American Dietetic Association*, vol. 94, pp. 57–64, 02 1994.

[43] J. Vesanto and E. Alhoniemi, “Clustering of the self-organizing map,” *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 11, pp. 586–600, 02 2000.

[44] P. Bholowalia and A. Kumar, “Ebk-means: A clustering technique based on elbow method and k-means in wsn,” *International Journal of Computer Applications*, vol. 105, no. 9, 2014.

[45] S. Na, L. Xumin, and G. Yong, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67.

[46] F. Murtagh and P. Legendre, “Ward’s hierarchical agglomerative clustering method: Which algorithms implement ward’s criterion?” *Journal of Classification*, vol. 31, no. 3, p. 274–295, Oct 2014. [Online]. Available: http://dx.doi.org/10.1007/s00357-014-9161-z

[47] S. C. Johnson, “Hierarchical clustering schemes,” *Psychometrika*, vol. 32, pp. 241–254, 1967.

[48] M. Forina, C. Armanino, and V. Raggio, “Clustering with dendrograms on interpretation variables,” *Analytica Chimica Acta*, vol. 454, no. 1, pp. 13–19, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267001015173

[49] V. Moosavi, S. Packmann, and I. Vallés, “Sompy: A python library for self organizing map (som),” 2014, gitHub.[Online]. Available: https://github. com/sevamoo/SOMPY.

A

Appendix

In this section of this thesis are some complementary tables or figures that are used to justify or present results shown in the body of the document. They'll be split into the different sections, so the reader can follow the structure and the order of the different methods used.

A.1 Parametric Values

The tables mentioned in Section 4.1 can be seen here:

Table A.1: Microbiological parameters

Parameter	Parametric Value	Unit
Escherichia coli (E. coli)	0	Number/100 ml
Enterococci	0	Number/100 ml

Table A.2: Chemical parameters

Parameter	Parametric Values	Unit
Acrylamide	0.10	μg/l
1,2 -dichloroethane	3.0	μg/l
Antimony	5.0	μg/l Sb
Arsenic	10.0	μg/l As
Benzene	1.0	μg/l
Benzo(a)pyrene	0.010	μg/l
Boron	1.0	mg/l B
Bromates	10.0	μg/l BrO_3
Cadmium	5.0	μg/l Cd
Chromium	50.0	μg/l Cr
Copper	2.0	mg/l Cu
Cyanides	50.0	μg/l CN
Epichlorohydrin	0.10	μg/l
Fluorides	1.5	mg/l F
Total Water Hardness	150-500	mg/l $CaCO_3$
Lead	10.0	μg/l Pb
Mercury	1.0	μg/l Hg
Nickel	20.0	μg/l Ni
Nitrates	50.0	mg/l NO_3
Nitrites	0.50	mg/l NO_2
PAHs	0.10	μg/l
Selenium	10.0	μg/l Se
Tetrachloroethene and Trichlorethene	10.0	μg/l
THMs	100.0	μg/l
Total pesticides	0.5	μg/l
Vinyl Chloride	0.50	μg/l

Table A.3: Indicative parameters

Parameter	Parametric Values	Unit
Aluminum	200.0	μg/l Al
Amonnium	0.50	mg/l NH_4
Calcium*	100.0	mg/l Ca
Chlorates	0.7	mg/l ClO_3
Chlorides	250.0	mg/l Cl
Chlorites	0.7	mg/l Cl_2
Clostridium perfringens	0.0	N/100 ml
Coliform bacteria	0.0	N/100 ml
Color	20.0	mg/l PtCo
Conductivity	2500	µS/cm a
Indicative Dose	0.10	mSv
Iron	200.0	µg/l Fe
Magnesium*	50.0	mg/l Mg
Manganese	50.0	µg/l Mn
Number of colonies at 22ºC*	100.0	N/ml at 22ºC
Number of colonies at 37ºC*	20.0	N/ml at 36ºC
Oxidability	5.0	mg/l O_2
pH	6.5 - 9.5	pH units
Radon	500.0	Bq/l
Residual disinfectant*	0.2 - 0.6	mg/l
Smell at 25ºC	3.0	dilution factor
Sodium	200.0	mg/l Na
Sulfates	250.0	mg/l SO4
Taste at 25ºC	3.0	dilution factor
TOC	without abnormal alteration	mg/l C
Total Water Hardness*	150 - 500	mg/l $CaCO_3$
Tritium	100.0	Bq/l
Turbidity	4.0	UNT

PAHs = Polycyclic Aromatic Hydrocarbons

THMs = Trihalomethanes

TOC = Total Organic Carbon

* = These are only recommended parametric values. In case of a range of values, it is recommended that the parameters' values are inside that interval.

A.2 Barreiro

A.2.1 Statistical Analysis

Regarding the Barreiro dataset, Subsection 6.1.1 referring to the statistical analysis, here are represented the boxplots of the parameters studied by years and their respective parametric or recommended value is marked in a red dotted line. These are illustrated from Figure A.1 to Figure A.12.

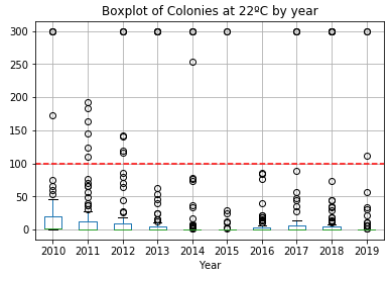

Figure A.1: Evolution of the number of colonies at 22ºC in the Barreiro Overview data set

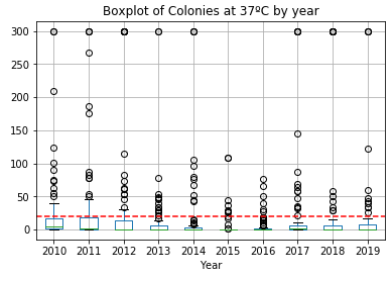

Figure A.2: Evolution of the number of colonies at 37ºC in the Barreiro Overview data set

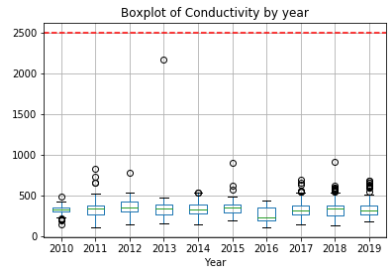

Figure A.3: Evolution of the conductivity in the Barreiro Overview data set

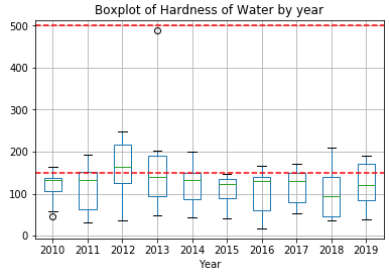

Figure A.4: Evolution of the hardness in the Barreiro Overview data set

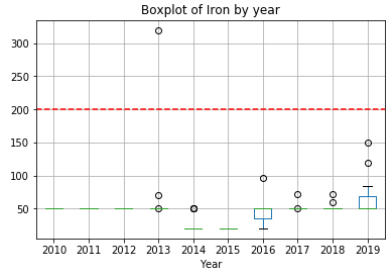

Figure A.5: Evolution of the iron in the Barreiro Overview data set

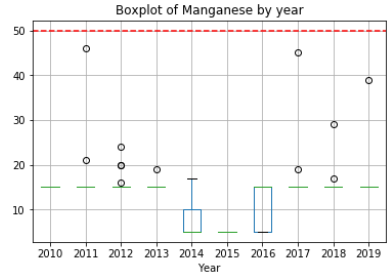

Figure A.6: Evolution of the manganese in the Barreiro Overview data set

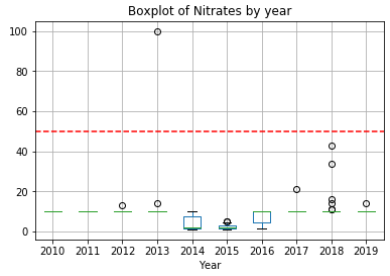

Figure A.7: Evolution of the nitrates in the Barreiro Overview data set

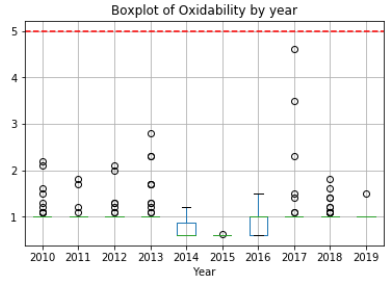

Figure A.8: Evolution of the oxidability in the Barreiro Overview data set

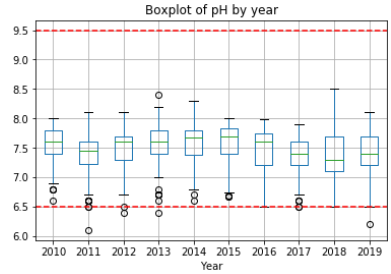

Figure A.9: Evolution of the pH in the Barreiro Overview data set

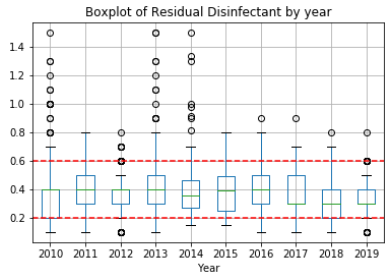

Figure A.10: Evolution of the residual disinfectant in the Barreiro Overview data set

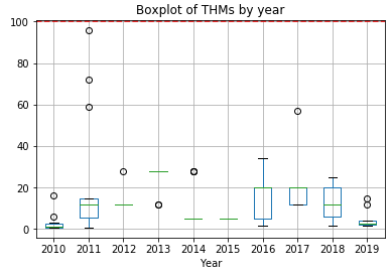

Figure A.11: Evolution of the THMs in the Barreiro Overview data set

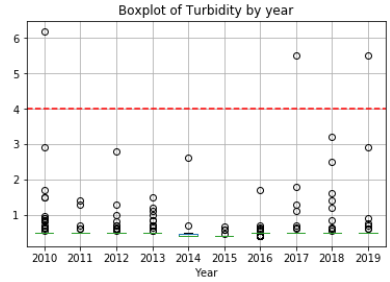

Figure A.12: Evolution of the turbidity in the Barreiro Overview data set

A.2.2 SOM analysis

Regarding the SOMs section, the hits map is represented in Figure A.13.

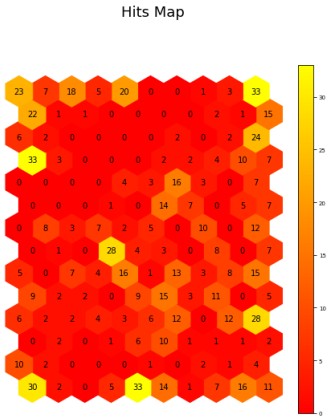

Figure A.13: Hits map of the Barreiro Overview data set

A.2.3 Cluster Analysis

The boxplots of the parameters in function of the clusters in the Barreiro data set are displayed in the following part. These are represented from Figure A.14 to A.22.

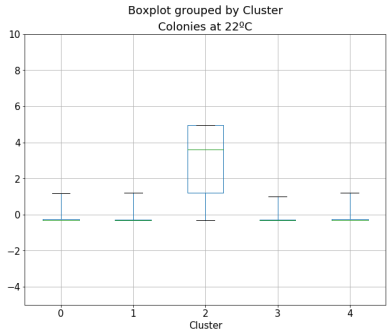

Figure A.14: Boxplot of the number of colonies at 22ºC in the Barreiro Overview filtered data set by cluster

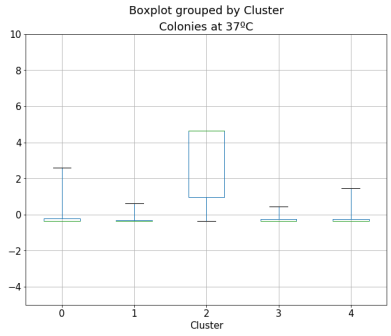

Figure A.15: Boxplot of the number of colonies at 37ºC in the Barreiro Overview filtered data set by cluster

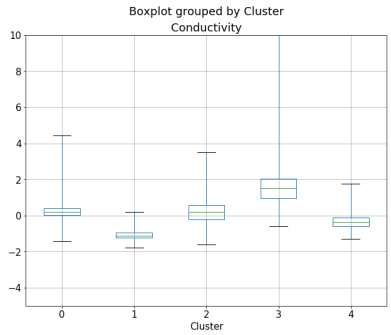

Figure A.16: Boxplot of the conductivity in the Barreiro Overview filtered data set by cluster

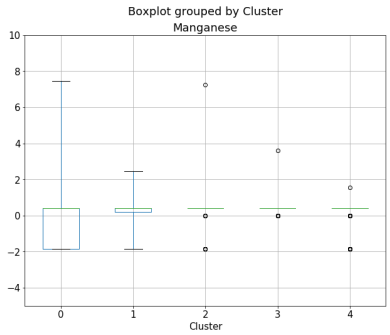

Figure A.17: Boxplot of the manganese in the Barreiro Overview filtered data set by cluster

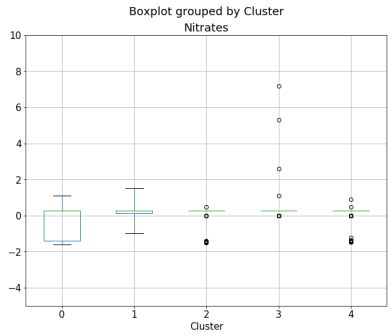

Figure A.18: Boxplot of the nitrates in the Barreiro Overview filtered data set by cluster

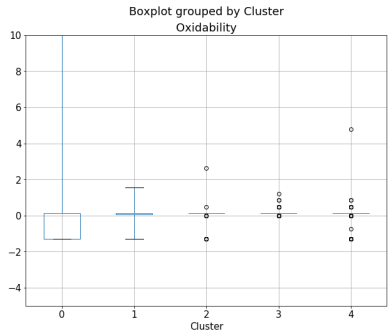

Figure A.19: Boxplot of the oxidability in the Barreiro Overview filtered data set by cluster

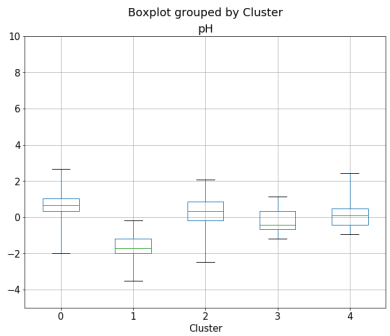

Figure A.20: Boxplot of the pH in the Barreiro Overview filtered data set by cluster

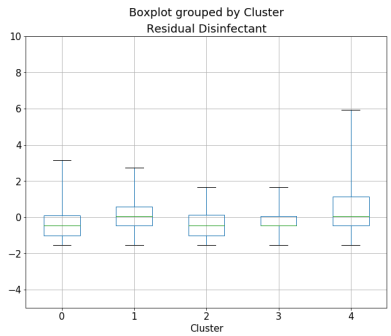

Figure A.21: Boxplot of the residual disinfectant in the Barreiro Overview filtered data set by cluster

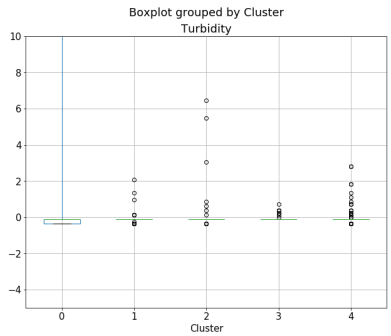

Figure A.22: Boxplot of the turbidity in the Barreiro Overview filtered data set by cluster

A.3 Beja

A.3.1 Statistical Analysis

Regarding the Beja dataset, Subsection 6.2.1 referring to the statistical analysis, here are represented the boxplots of the parameters studied by years and their respective parametric or recommended value is marked in a red dotted line. These are illustrated from Figure A.23 to Figure A.36.

A.3.2 SOM analysis

Regarding the SOMs section of the Beja data set, the hits map is represented in Figure A.37.

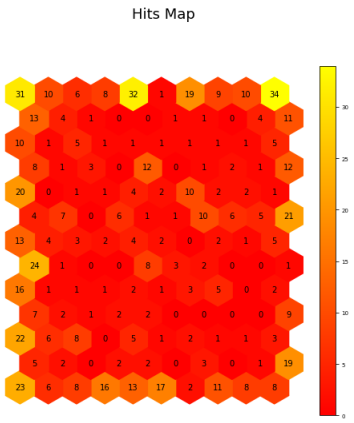

Figure A.37: Hits map of the Beja Overview data set

A.3.3 Cluster Analysis

Regarding the clusters analysis in the Beja data set, the boxplots of the parameters in function of the clusters are displayed in the following part. These are represented from Figure A.38 to A.46.

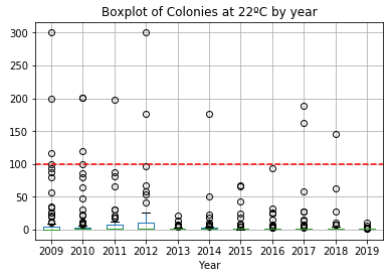

Figure A.23: Evolution of the number of colonies at 22ºC in the Beja Overview data set

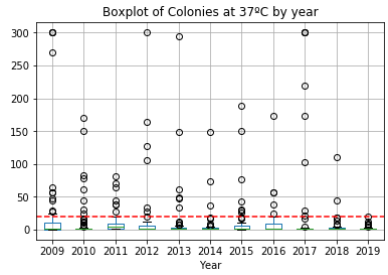

Figure A.24: Evolution of the number of colonies at 37ºC in the Beja Overview data set

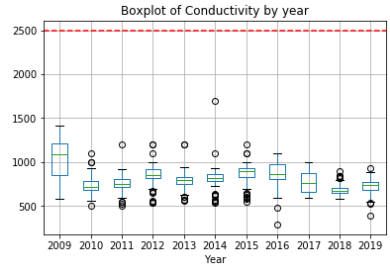

Figure A.25: Evolution of the conductivity in the Beja Overview data set

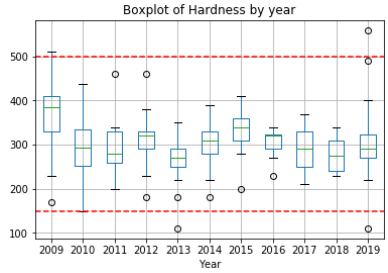

Figure A.26: Evolution of the hardness in the Beja Overview data set

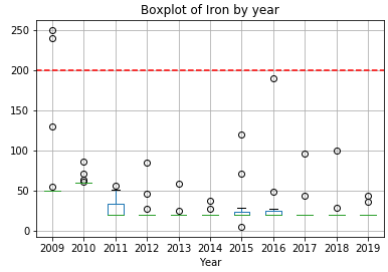

Figure A.27: Evolution of the iron in the Beja Overview data set

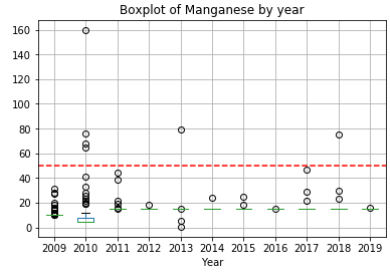

Figure A.28: Evolution of the manganese in the Beja Overview data set

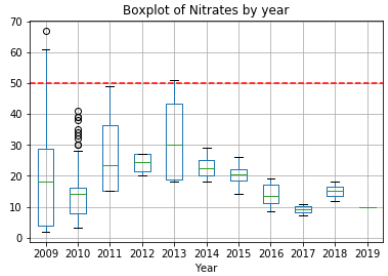

Figure A.29: Evolution of the nitrates in the Beja Overview data set

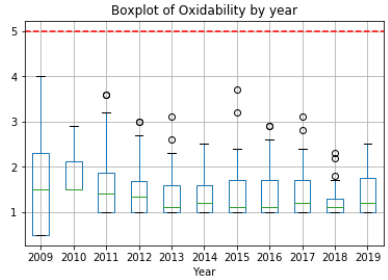

Figure A.30: Evolution of the oxidability in the Beja Overview data set

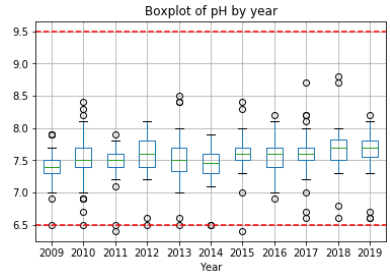

Figure A.31: Evolution of the pH in the Beja Overview data set

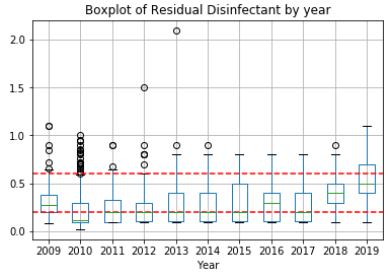

Figure A.32: Evolution of the residual disinfectant in the Beja Overview data set

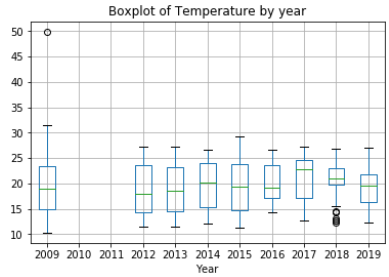

Figure A.33: Evolution of the temperature in the Beja Overview data set

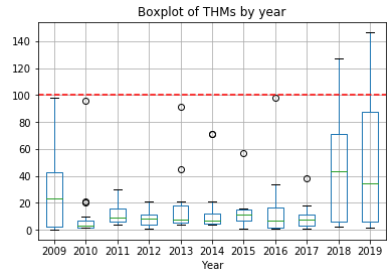

Figure A.34: Evolution of the THMs in the Beja Overview data set

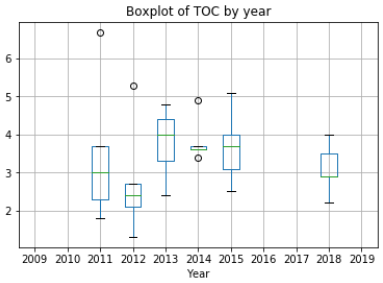

Figure A.35: Evolution of the TOC in the Beja Overview data set

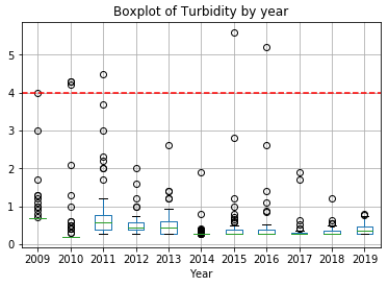

Figure A.36: Evolution of the turbidity in the Beja Overview data set

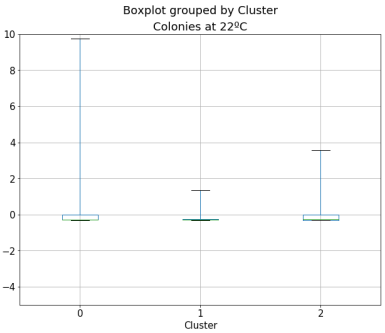

Figure A.38: Boxplot of the number of colonies at 22ºC in the Beja Overview filtered data set by cluster

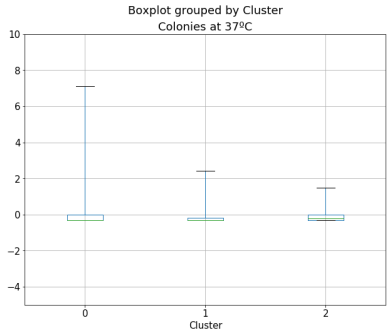

Figure A.39: Boxplot of the number of colonies at 37ºC in the Beja Overview filtered data set by cluster

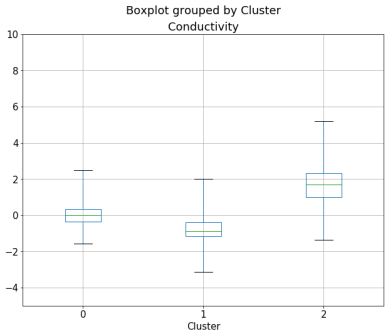

Figure A.40: Boxplot of the conductivity in the Beja Overview filtered data set by cluster

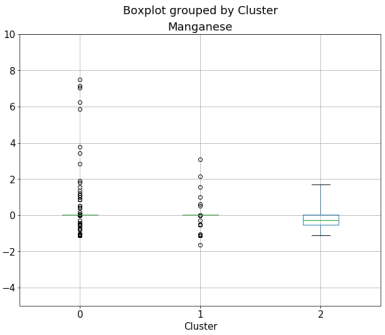

Figure A.41: Boxplot of the manganese in the Beja Overview filtered data set by cluster

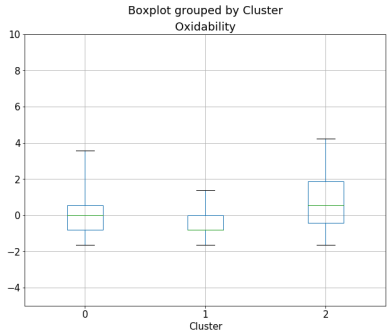

Figure A.42: Boxplot of the oxidability in the Beja Overview filtered data set by cluster

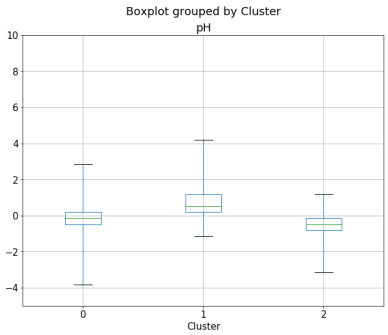

Figure A.43: Boxplot of the pH in the Beja Overview filtered data set by cluster

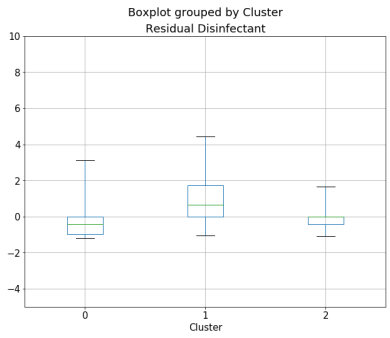

Figure A.44: Boxplot of the residual disinfectant in the Beja Overview filtered data set by cluster

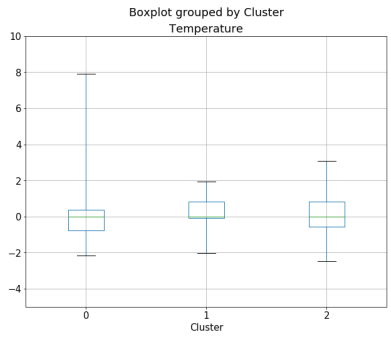

Figure A.45: Boxplot of the temperature in the Beja Overview filtered data set by cluster

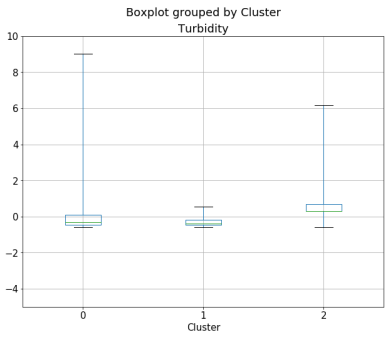

Figure A.46: Boxplot of the turbidity in the Beja Overview filtered data set by cluster

A.4 Infraquinta

A.4.1 Statistical Analysis

Concerning the Infraquinta dataset, Subsection 6.3.1 referring to the statistical analysis, here are represented the boxplots of the parameters studied by years and their respective parametric or recommended value is marked in a red dotted line. These are illustrated from Figure A.47 to Figure A.57.

A.4.2 SOM analysis

Regarding the SOMs section, the hits map is represented in Figure A.58.

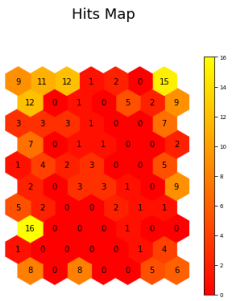

Figure A.58: Hits map of the Infraquinta Overview data set

A.4.3 Cluster Analysis

Regarding the clusters analysis in the Infraquinta data set, the boxplots of the parameters in function of the clusters are displayed in the following part. These are represented from Figure A.59 to A.66.

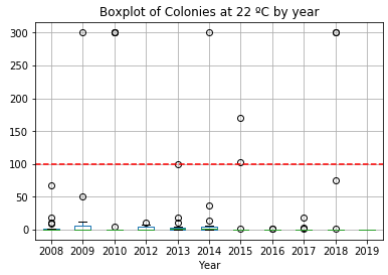

Figure A.47: Evolution of the number of colonies at 22℃ in the Infraquinta Overview data set

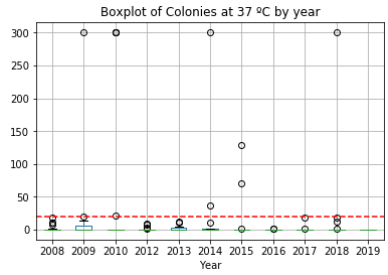

Figure A.48: Evolution of the number of colonies at 37℃ in the Infraquinta Overview data set

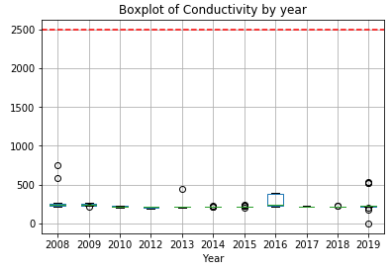

Figure A.49: Evolution of the conductivity in the Infraquinta Overview data set

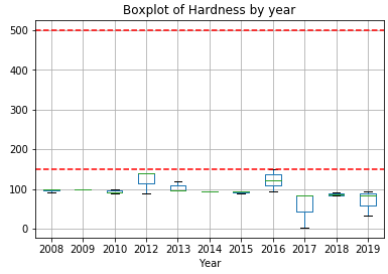

Figure A.50: Evolution of the hardness in the Infraquinta Overview data set

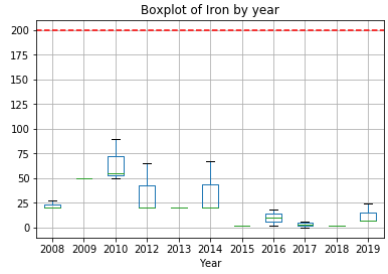

Figure A.51: Evolution of the iron in the Infraquinta Overview data set

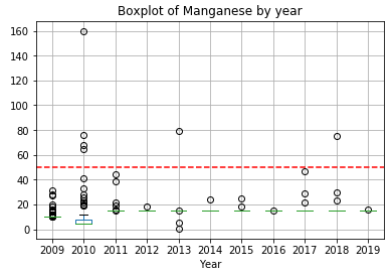

Figure A.52: Evolution of the manganese in the Infraquinta Overview data set

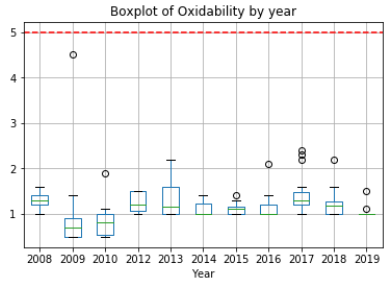

Figure A.53: Evolution of the oxidability in the Infraquinta Overview data set

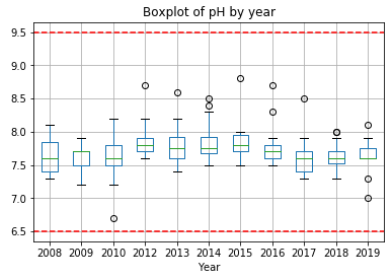

Figure A.54: Evolution of the pH in the Infraquinta Overview data set

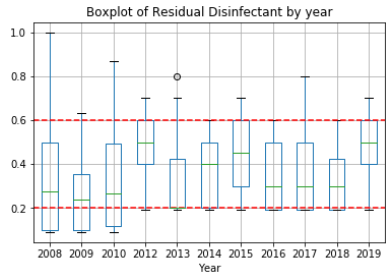

Figure A.55: Evolution of the residual disinfectant in the Infraquinta Overview data set

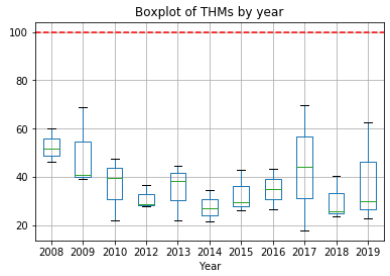

Figure A.56: Evolution of the THMs in the Infraquinta Overview data set

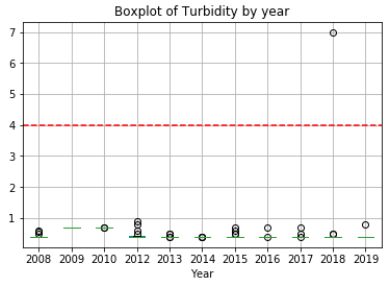

Figure A.57: Evolution of the temperature in the Infraquinta Overview data set

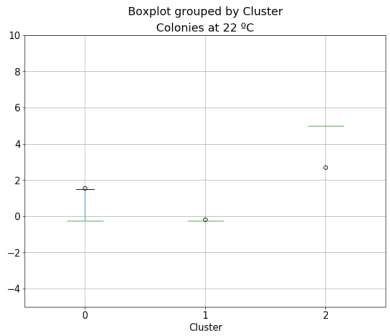

Figure A.59: Boxplot of the number of colonies at 22℃ in the Infraquinta Overview filtered data set by cluster

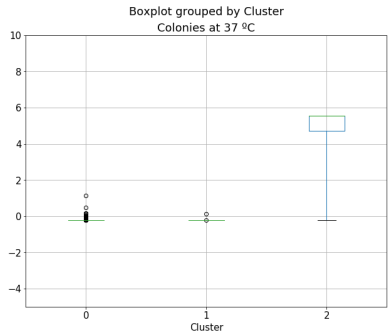

Figure A.60: Boxplot of the number of colonies at 37℃ in the Infraquinta Overview filtered data set by cluster

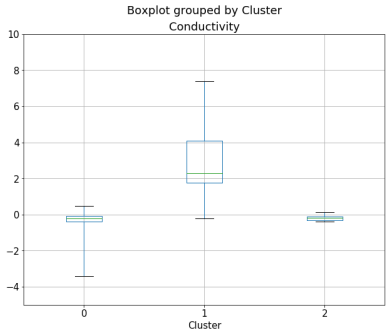

Figure A.61: Boxplot of the conductivity in the Infraquinta Overview filtered data set by cluster

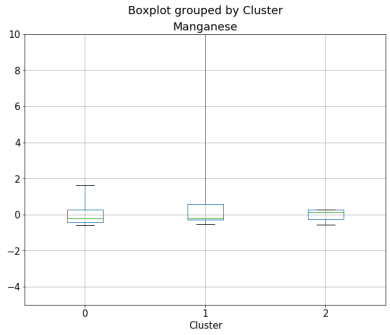

Figure A.62: Boxplot of the manganese in the Infraquinta Overview filtered data set by cluster

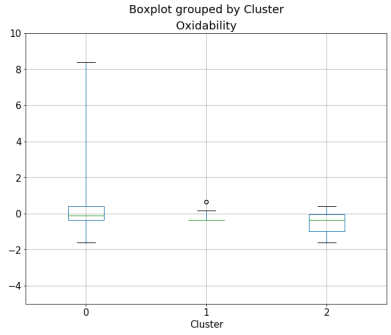

Figure A.63: Boxplot of the oxidability in the Infraquinta Overview filtered data set by cluster

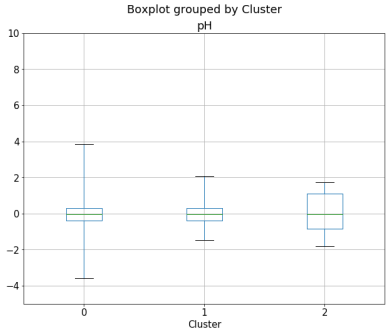

Figure A.64: Boxplot of the pH in the Infraquinta Overview filtered data set by cluster

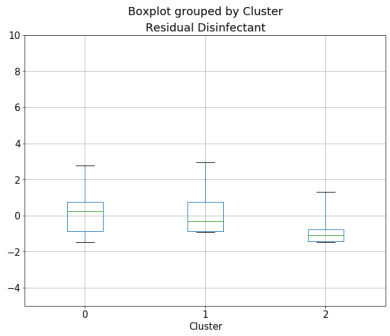

Figure A.65: Boxplot of the residual disinfectant in the Infraquinta Overview filtered data set by cluster

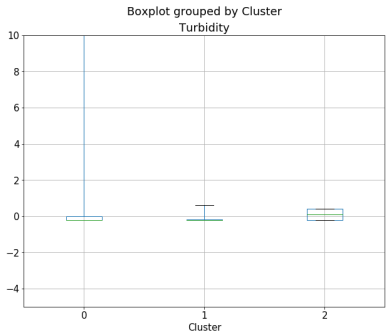

Figure A.66: Boxplot of the turbidity in the Infraquinta Overview filtered data set by cluster