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

Rodrigo M. Ribeiro

rodrigomendesribeiro@tecnico.ulisboa.pt

Abstract—Over the years water managing entities 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 and over the minimum values required. This paper investigates the water quality data gathered by a water distribution managing entity over 10 years in Barreiro, Portugal. With this information, one 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). For this, 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 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 (K-Means and Hierarchical Clustering) were employed. As to the WQIs, 2 of them are adaptations from already existing work and the 3rd one is a completely new approach using differences of the actual concentration of the parameter to its parametric value. Among many results, it was concluded that there is a strong relation between number of colonies at 22ºC and 37ºC, the total water hardness between the conductivity and pH. These relations featured in the remaining methods used. It was also found that, as expected, a big majority of the parameters were found within the parametric values, resulting in a very good or excellent water quality for the 3 indexes it was evaluated.

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

I. Introduction

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. Only 0.08% of all the fresh water on Earth is actually used and exploited by humankind [1] [2]. 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. To answer this problem, among many other around the world, the WISDom project (Water Intelligence Systems Data project) [3] was created. 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. Besides the interest of making the best usage possible for this asset, it is also important to guarantee a quality water for the customers, ensuring its satisfaction and public health at the same time. Monitoring the parameters that establish the water quality is vital to safeguard everyone's safety and contentment.

Over the last few years, companies and managing entities in charge of water networks have collected data for no other purpose but to report to the government to make sure the water being delivered is within the parametric values. Hence, in conformity with the legislation. This results in an enormous quantity of data with no actual treatment. With this immense information gathered and with the advancements in the areas of machine learning it has been possible to develop several techniques that allow the researchers and consequently the water utilities to uncover new revealing information regarding the network and relations between the parameters that were yet to be assessed. Lately, there has been an increase number of studies in the water quality studies utilizing a class of Artificial Neural Networks (ANN) [4] [5] [6] called Self-Organizing Maps (SOMs) [7]. These allow a better visualization of the data being analyzed, achieved by reducing its dimensions to a 2D plane. Along with this kind of study, it is usual to observe other methods such as correlations, Principal Component Analysis (PCA) and clustering techniques [8] [9]. SOMs are considered a more advanced technique than these ones, since they are used to solve multivariate problems, while the rest only solves them in a linear fashion.

Despite this analysis, it is unusual to see it followed by a WQI study. For this matter, it was decided to include this part in the same paper as the aforementioned analysis. Regarding the WQIs, most of the works ignore the microbiological parameters present in the water, since they represent a difficult parameter to insert and take into account in the indexes [10] [11] [12]. This paper offers alternative indexes that take this fact in consideration. Thus, giving a more realistic and complete view of the water quality in the water network.

The aim of this paper is to analyze the data provided by the C.M. Barreiro from 2010 to 2019. This data consists of the values of the concentrations obtained for each parameter evaluated in the mentioned period.

This paper proposes to demonstrate the utility of the application of SOMs as an effective method to visualize and simplify complex multivariate problems such as the one present in the water quality data, while preserving the structure of the initial data. The unsupervised analysis is then followed. It includes a Principal Component Analysis and two clustering techniques represented in the Kohonen SOMs produced. Afterwards, the WQIs evaluation of the new proposed indexes is performed.

II. Materials and Methods

a. Feature Selection

1. Data selection and data cleaning

Upon receiving the data from the managing entity, it was noticeable that not all the parameters are of interest. So, some of them had to be removed according to their importance and relevance. 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.º 306/2007, of August 27th, with the changes introduced by Decreto-Lei n.º 152/2017, of December 7th.

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.

Another reason for exclusion can be an insufficient number of observations,which didn't make sense to study in this case. Another possibility is if there is just not enough differentvalues for the parameter in question A new filter had to applied to decrease the number of parameters because not all parameters have the same importance or relevance to study. Another reason for exclusion can be an insufficient number of observations, which did not make sense to study in this case. Another possibility is if there is just not enough different values for the parameter in question. For these reasons, only 12 parameters were selected: the number of colonies at 22ºC and at 37ºC, conductivity, hardness of water or total water hardness, iron, manganese, nitrates, oxidability, pH, residual disinfectant, trihalometanos (THMs) and turbidity.

The data cleaning process is a key process, because the data received from the managing entity had errors such as the same parameters written in different ways or with orthographic errors, not all variables received were of interest and were immediately removed, and some results of the parameters came with symbols attached to them that had to be interpreted (and if necessary make changes to other variables) and then removed as well to keep the result as a single and normal value.

2. Outlier detection

The outlier detection is important because the results for all the parameters studied need to make sense so they can be relevant for the study of the distribution network. 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.

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.

b. 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. For the relevant parameters, a further investigation is conducted. In this work, a correlation above the absolute value of 0.6 is considered of interest. In this examination, the focus is to demonstrate how the the evolution of the correlation between pairs of parameters occurs and how significant it is.

c. SOM Analysis

In this paper SOMs are utilized both for analyzing the correlation between the different water parameters in a multivariate way as well as for clustering applications. They represent a type of unsupervised ANN that utilizes competitive learning that performs a dimensionality reduction of the data into a 2D plane preserving all of its original topographic properties. This technique was first introduced by the finish professor Teuvo Kohonen in the 1980s and it has been widely used in these kind of studies for its capability to present complex data in a simplified and clear way - the 2D plane. This approach has the advantage of representing high dimension data in a clearer way that provides simpler visual comprehension and interpretation of multi-dimensional and complex data sets, revealing non-linear properties not easily detected. The SOM learning process generates a many-to-one mapping between the input data and map units. The map units are arranged in a 2D lattice and each is associated with a weight vector. Considering each observation a vector $x_1, x_2 ..., x_L$ of dimension L. Since it is an iterative process, 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 parameters evaluated. 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:

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

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.

$$w_{ij}(t+1) = w_{ij}(t) + R(t)\eta(t)(x_j^k - w_{ij}(t)) \quad (2)$$

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.

Since the optimal number of nodes is near $5\sqrt{N}$, where N is the total number of observations, the map size is defined accordingly. In this work, 10 models are trained and the respective topographic and quantization errors are retrieved. The model with the lowest topographic error is chosen, since the quantization one does not show much variation. 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.

d. Other Unsupervised methods

1. Principal Component Analysis (PCA)

PCA is an efficient tool to explain the variance of a large data set with a short number of uncorrelated principal components (PC). Multiplying the parameters with the eigenvector results in the PCs. These can provide information of the most important parameters that describe the data set allowing data reduction with minimum loss of original data. 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. A correlation above the absolute value 0.5 is considered important

2. KMeans

Kmeans is a clustering technique that tries to partition the data set into K pre-defined groups. Data points are assigned to a certain cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The number of K clusters is previously determined by an elbow plot. This plot is accomplished using the sum of squared distance (SSE) between data points and their assigned clusters' centroids.

3. Hierarchical Clustering

The hierarchical clustering technique used was an agglomerative method using Ward linkage. This method minimizes the total within-cluster variance. To determine the optimal number of clusters, a manual inspection of the corresponding dendogram was done. It was always in concordance with the KMeans elbow plot, which helped to verify the credibility of both methods.

4. Clustering Analysis

In order to perform the analysis of the clusters. 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:

$$\tilde{x} = \frac{x - \bar{x}}{\mathrm{s}} \quad (3)$$

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.

e. Water Quality Index (WQI)

Water quality index (WQI) is defined as a rating reflecting the composite influence of different water quality parameters. For this work, three new indexes are proposed (indexes A, B and C). The first two are adaptations from previous works in the field and the last one is a water quality index based on the differences between the parametric values and the actual concentrations of the parameters. In all three of them, there was a special attention dedicated to use as many parameters as possible as well as the inclusion of microbiological parameters in the calculation of the indexes.

1. Index A

Considering C_i the concentration for the parameter i, C_0 the ideal value for the parameters (7 for the pH and 0 for all the others), S_i the standard value for the parameter i and N the total number of observations, the index is built in the following way:

- Quality index (Qi): $\frac{C_i - C_0}{S_i - C_0} \times 100$
- Weight index (Wi): $\frac{1}{S_i}$

- WQI: $\frac{\sum^N W_i Q_i}{\sum^N W_i}$

Regarding the scale to quantify the actual quality of the water it is considered the same used in works such as Akter et al. [11] and Yisa et al.[10], since the WQI used is the very similar. This is represented in Table 1.

Table 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), coliform bacteria and enterococci - 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.

2. Index B

The second index proposed, index B, 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 and represents the only difference from the first index. The flag system is the same as in index A. 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°C, conductivity, color, manganese, nitrates, number of colonies at 22°C and 37°C, oxidability, taste at 25°C, 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, 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 1).

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

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. 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. Another 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. 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.

This index can be summed in a few steps:

1. Normalize the data and parametric values (values and ranges) using min-max normalization;
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.

The classification of the results obtained can be seen in Table 2.

Table 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

III. Case Study

The case study is the data set provided by the C.M. Barreiro regarding water quality in its network/distribution. Barreiro is a portuguese city located in Setúbal district. This network provides water to about 80000 inhabitants. The data received from this managing entity consists of 2192 observations with 47 different parameters after the outlier removal process. Each observation consists of a pair sampling site and date. So, with this format each variable is a different parameter. It can happen that for one location on a specific date, several parameters weren't tested resulting in null (NaNs) values. However, if a certain parameter was tested, then the obtained concentration is presented.

IV. Results and Discussion

a. Statistical Analysis

Of all the parameters present in the original data set, not all have the same importance, as explained previously. So, from now on only 12 parameters are analyzed in more depth: the number of colonies at 22°C and at 37°C, conductivity, hardness, iron, manganese, nitrates, oxidability, pH, residual disinfectant, THMs and turbidity. 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. These results are found in Table 3.

Table 3: Descriptive analysis of the parameters - Barreiro overall view

Parameter	$\bar{x}$	s	Min	Max	PV	no. Obs.
Colonies at 22°C*	17.71	56.80	0.0	301.0	100	783
Colonies at 37°C*	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°C and 37°C 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.

b. 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. 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 1.

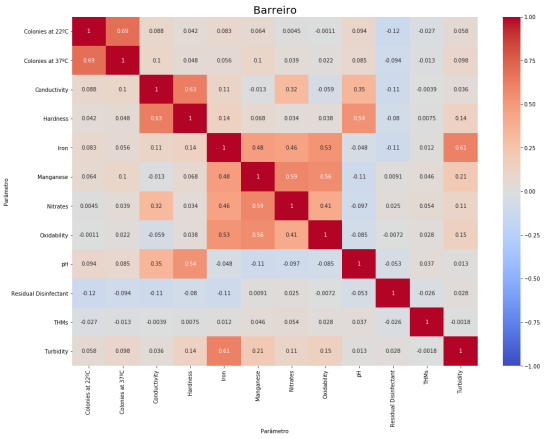


Fig. 1: Correlation Matrix of the Barreiro Overview filtered data set

The association between the number of colonies at 22°C and 37°C 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°C, 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 2.

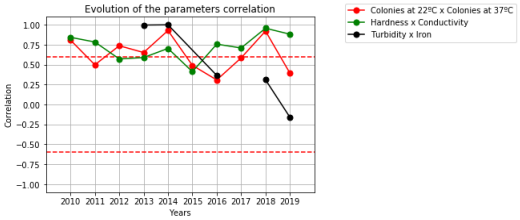


Fig. 2: Correlation of parameters throughout time in Barreiro Overview filtered data set

Two pairs of parameters seem to show a very stable high correlation such as the number of colonies at 22°C and at 37°C (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.

c. SOMs

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 3.

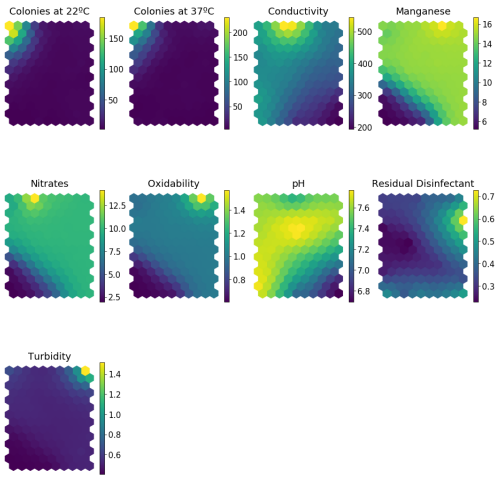


Fig. 3: SOM of Barreiro Overview filtered data set

Once more, as expected, the number of colonies at 22°C and 37°C 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.

d. PCA

For this data set, 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 4, the correlations of the different principal components with the parameters in the original data are represented. These are often called factor loadings.

TABLE 4: BARREIRO OVERVIEW FILTERED DATA SET - PRINCIPAL COMPONENTS ANALYSIS

Parameter	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Colonies at 22°C	0.18	0.59	-0.32	-0.01	-0.15	-0.01
Colonies at 37°C	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°C and 37°C. This is an indicator that this component increases with increasing number of colonies at 22°C and 37°C. 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.

e. 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 4 and 5.

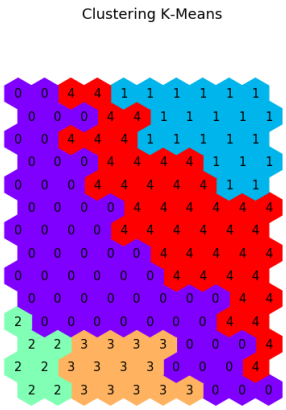


Fig. 4: Barreiro Overview filtered data set - K-Means in SOM

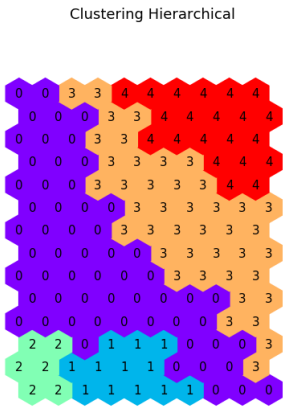


Fig. 5: Barreiro Overview filtered data set - hierarchical clustering in SOM

From the two Figures (4 and 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 Figure. The clusters are represented in Figure 6.

The interpretation of the clusters obtained starts with cluster 0. Here, 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°C and at 37°C 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.

f. 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 5.

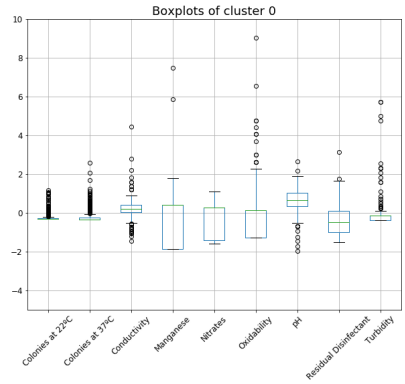
TABLE 5: RATING CLASSIFICATION FOR INDEX A, B AND C FOR THE BARREIRO DATA SET

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

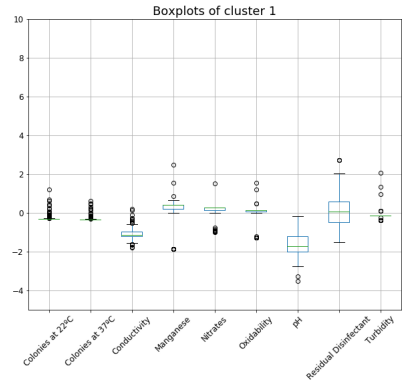
Values marked with a "-" mean that for that specific rating, the index does not have a classification defined. As it is observable, the main quantity of the observations are rated as excellent water, which is a very good sign for the water utility. One noticeable fact from this index is that there are 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 5.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.

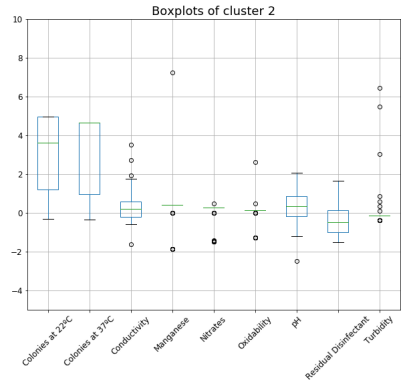
For index C, the results obtained are also displayed in Table 5. For this index, it is much more difficult to actually make the difference between the good, very good and excellent quality of the water. The high amount of observations being classified as good derives from the fact that a lot of observations 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



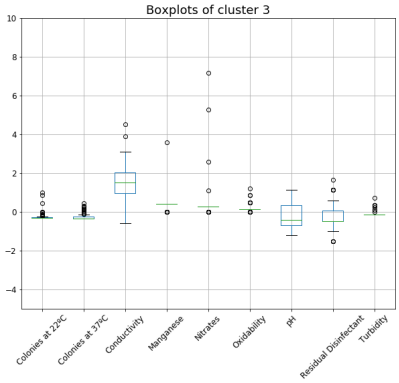
(a) Barreiro Overview filtered data set-Cluster 0



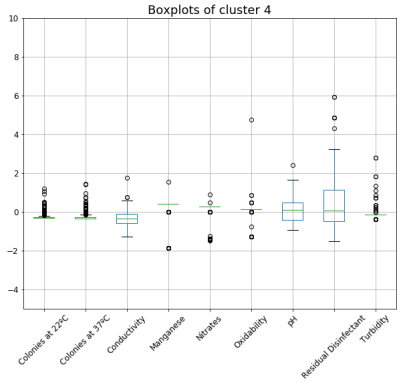
(b) Barreiro Overview filtered data set-Cluster 1



(c) Barreiro Overview filtered data set-Cluster 2



(d) Barreiro Overview filtered data set-Cluster 3



(e) Barreiro Overview filtered data set - Cluster 4

Fig. 6: Barreiro Overview filtered data set - Clusters

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.

V. Conclusion

a. Conclusions

This paper is based on a work that contained two more case studies. One regarding the water utility of C.M. Barreiro, a municipality in Setúbal district, Portugal, and another concerning EMAS Beja, which is a water utility that provides water to the inhabitants of Beja, the municipality capital of the Beja district in Portugal. With regard to the Barreiro data set analyzed in this work, the conclusions are displayed by Section.

Regarding the correlation analysis, The most notorious correlations found in Barreiro after a more detailed study were the relation between the number of colonies at 22°C and at 37°C and the relation between hardness and conductivity. These tend to show a steady correlation above the 0.6 threshold across almost every year. The relation between the number of colonies at 22°C and at 37°C was to be expected, since it makes sense from the microbiological sense. If there are microorganisms at 22°C, it becomes clear that an augment on this parameter would influence positively the presence of harmful microorganisms at 37°C.

As to the SOM analysis, 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.

Concerning the Principal Component Analysis (PCA), there are 6 principal components, since one PC often represents or correlates highly with one or two parameters. The main conclusions are already stated in Section IV Subsection d. Comparing to the other case studies analyzed, the principal component common in all the case studies is one component that correlates highly with the number of colonies at 22°C and at 37°C. One justification for this is the fact that these two parameters already correlate highly with each other.

Regarding the cluster obtained after performing the SOMs, some interesting results were found. The Barreiro data set had an optimal number of clusters of 5, mainly due to a high number of observations and the consequent increase of complexity in the data. The most noticeable cluster is one carrying high variability and high average and median levels for the number of colonies at 22°C and at 37°C. In this data set, cluster 3 showed a high variability and usually higher levels for the conductivity. The residual disinfectant also played a role when building the clusters, since cluster 4 showed high variability for this parameter.

As to the three new proposed indexes, 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. Thus, for indexes A and B the quality of the water is ranked as excellent. The classification given by index C ranks the water quality as very good.

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, 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.

b. 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.

REFERENCES

[1] 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

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

[3] https://wisdom.ips.pt/.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.