AgriEngineering

MDPI

Article

A Data-Driven Method for Water Quality Analysis and Prediction for Localized Irrigation

Roberto Fray da Silva [1,2,*], Marcos Roberto Benso [2,3], Fernando Elias Corrêa [2], Tamara Guindo Messias [4], Fernando Campos Mendonça [1], Patrícia Angelica Alves Marques [2,5], Sergio Nascimento Duarte [1], Eduardo Mario Mendiondo [2,3], Alexandre Cláudio Botazzo Delbem [2,6], Antonio Mauro Saraiva [2,7,8]

[1] Biosystems Engineering Department, ESALQ, University of Sao Paulo, Av. Pádua Dias, 11, Piracicaba 13418-900, SP, Brazil; fernando.mendonca@usp.br (F.C.M.); snduarte@usp.br (S.N.D.)
[2] Center for Artificial Intelligence—C4AI, University of Sao Paulo, Av. Prof. Lúcio Martins Rodrigues, 370-Butantã, São Paulo 05508-020, SP, Brazil; marcosbenso@alumni.usp.br (M.R.B.); fecorrea@usp.br (F.E.C.); paamarques@usp.br (P.A.A.M.); emm@sc.usp.br (E.M.M.); acbd@icmc.usp.br (A.C.B.D.); saraiva@usp.br (A.M.S.)
[3] TheWADILab, CEPED, EESC, University Sao Paulo, Av. Trabalhador Saocarlense, 400, São Carlos 13566-590, SP, Brazil
[4] ESALQ, University of Sao Paulo, Av. Pádua Dias, 11, Piracicaba 13418-900, SP, Brazil; tamessias@gmail.com
[5] PPGESA, Biosystems Engineering Department, ESALQ, University of Sao Paulo, Av. Pádua Dias, 11, Piracicaba 13418-900, SP, Brazil
[6] Institute of Mathematics and Computer Sciences, University of Sao Paulo, Av. Trab. São Carlense, 400-Centro, São Carlos 13566-590, SP, Brazil
[7] Polytechnic School, University of Sao Paulo, Av. Prof. Luciano Gualberto, 380-Butantã, São Paulo 05508-010, SP, Brazil
[8] Institute of Advanced Studies, University of Sao Paulo, R. da Praça do Relógio, 109-Conj. Res. Butanta, São Paulo 05508-050, SP, Brazil
* Correspondence: roberto.fray.silva@usp.br

check for updates

Citation: da Silva, R.F.; Benso, M.R.; Corrêa, F.E.; Messias, T.G.; Mendonça, F.C.; Marques, P.A.A.; Duarte, S.N.; Mendiondo, E.M.; Delbem A.C.B.; Saraiva, A.M. A Data-Driven Method for Water Quality Analysis and Prediction for Localized Irrigation. *AgriEngineering* **2024**, *6*, 1771–1793. https://doi.org/10.3390/agriengineering6020103

Academic Editors: José Manuel Monteiro Gonçalves, Giovanni Rallo

Received: 15 March 2024
Revised: 24 May 2024
Accepted: 4 June 2024
Published: 18 June 2024

Abstract: Several factors contribute to the increase in irrigation demand: population growth, demand for higher value-added products, and the impacts of climate change, among others. High-quality water is essential for irrigation, so knowledge of water quality is critical. Additionally, water use in agriculture has been increasing in the last decades. Lack of water quality can cause drip clog, a lack of application uniformity, cross-contamination, and direct and indirect impacts on plants and soil. Currently, there is a need for more automated methods for evaluating and monitoring water quality for irrigation purposes, considering different aspects, from impacts on soil to impacts on irrigation systems. This work proposes a data-driven method to address this gap and implemented it in a case study in the PCJ river basin in Brazil. The methodology contains nine components and considers the main steps of the data lifecycle and the traditional machine learning workflow, allowing for automated knowledge extraction and providing important information for improving decision making. The case study illustrates the use of the methodology, highlighting its main advantages and challenges. Clustering different scenarios in three hydrological years (high, average, and lower streamflows) and considering different inputs (soil-related metrics, irrigation system-related metrics, and all metrics) helped generate new insights into the area that would not be easily obtained using traditional methods.

Keywords: clustering; case study; data-driven methodology; unsupervised learning; water monitoring; water quality

1. Introduction

Exponential population growth has led to an increasing demand for food, which has resulted in a need to improve farm productivity. However, the availability of arable land and quality water has been reduced due to several factors, such as climate change, erosion, water pollution by different sources, and land cover changes [1,2].

Research about the mechanisms influencing surface water quality is paramount, especially considering the extensive amount of data amassed in specific studies. Pollution from natural origins, intensive agricultural activities, and rapid urban expansion substantially stress water resources [3].

Agriculture is the leading water consumer in the world, accounting for 87% of global water consumption and 60% of all freshwater capture [4]. In Brazil, approximately 60% of the freshwater collected is used for agriculture [5]. However, the water quality can vary considerably in the different regions and seasons. Low-quality irrigation water can impact plant growth, soil quality, and clog irrigation systems. This is a significant problem for localized irrigation.

Therefore, monitoring irrigation water quality is critical for sustainably and adequately managing irrigated agriculture. The global shortage of fresh water is a serious issue that will worsen with increasing demand and the effects of climate change. Precision Agriculture and Smart Irrigation could be essential solutions for addressing these issues [6].

It is fundamental to monitor water quality, especially to measure its impacts on agriculture, ecosystems, and water resources [7–9]. Several indices can be used to provide a clear overview of the state and comprehensively evaluate surface water quality [3]. However, there are many questions about which parameters are available, the quality and frequency of obtaining these parameters, and the gaps in the available data. Answering these questions is essential to help improve policy-making decisions.

Additionally, the better monitoring of water quality can help farmers and governments allocate resources and technologies in agricultural areas, increasing the resilience of water resources to anthropogenic impacts. Nevertheless, the traditional method of evaluating water purity for both drinking and irrigation purposes is characterized by its high cost, time-intensive nature and substantial demand for personnel resources [3]. In this context, automated methods for evaluating and monitoring water quality for irrigation are essential for dealing with large volumes of data. Additionally, it is essential to conduct these evaluations considering different aspects, including soil impacts on irrigation systems.

Machine learning (ML) algorithms can help build such automated models by capturing linear and nonlinear relationships between hydrological systems and using parameters in heterogeneous areas subject to different land management systems and varying anthropogenic impacts. Traditionally, ML is divided into three main areas: supervised, unsupervised, and reinforcement learning.

According to James et al. [10], unsupervised learning is a set of techniques for exploring data and identifying relationships between parameters and the data structure. It is used when the specific labels of the data points are unknown ((limiting the opportunities for supervised learning), typically with complex, real-world problems. Usually, unsupervised learning is divided into two main sets of techniques: clustering and dimensionality reduction.

Clustering methods are widely used to explore data structures in such contexts. One of the main objectives of using these methods is identifying groups of parameters (also called features) or data points according to predefined criteria. Typically, a criterion is the distance between the data points in an n-dimensional space [11]. Clustering is also widely used for outlier detection.

There is an important gap in the literature concerning automatic methods for evaluating water quality for irrigation purposes, considering indices or metrics that belong to different dimensions (such as soil-related and irrigation system-related). As described in this section, artificial intelligence (AI) techniques, especially unsupervised learning, can provide important tools to improve knowledge extraction and decision making for evaluating water quality for irrigation. This could reduce mistakes that cause considerable problems in localized irrigation, such as drip clogging.

Recent developments in data-driven models present the opportunity to learn from historical data without any predefined relationship among the parameters [12]. Despite the recent developments of data-driven methods in the literature, significant challenges

remain, including increased uncertainty caused by the impact of hydrological extremes, such as droughts and floods, on water quality.

The central research gap addressed in this work is the need for more automated methods for evaluating and monitoring water quality for irrigation purposes, considering different aspects, from impacts on the soil to impacts on irrigation systems, that can be used in other regions and with other variables or quality indices as inputs. Additionally, the results of the proposed framework should provide vital information to improve decision making.

Therefore, this work has two main objectives: (i) propose a data-driven method to analyze and predict water quality for irrigation purposes; and (ii) conduct an in-depth case study considering the proposed method for cities with a high water demand for irrigation use. The case study encompasses two quality indicators for soil-related purposes (electrical conductivity and pH) and four for irrigation system-related purposes (total iron, hardness, biochemical oxygen demand, and thermotolerant coliforms). It is essential to observe that the code developed could be adapted to analyze water quality for other purposes, such as human consumption, and used with other parameters, quality indices, regions, and river basins.

To address these objectives, this work aimed to answer two research questions (RQs):

- RQ1: What components should be considered to develop a data-driven water quality analysis and monitoring methodology for irrigation-related purposes?
- RQ2: How did the water quality vary in the studied areas for the three hydrological years considered (flood, drought, and average year), considering indices related to soil and irrigation systems?

The main contributions of this study are (i) propose and implement a data-driven methodology to analyze and monitor water quality for irrigation purposes; (ii) apply an unsupervised learning technique on real data from a complex problem with economic, environmental, and social impacts; and (iii) have two domain experts conduct an in-depth evaluation of the results.

This rest of this paper is organized into the following sections: Section 2 describes the theoretical foundations, encompassing relevant descriptions of water quality metrics for irrigation and the use of AI for water quality evaluation; Section 3 describes the data-driven methodology proposed; Section 4 details the study case and the main results of applying the proposed method to its data; Section 5 encompasses a discussion of the main impacts of the results for the case study and of the use of the methodology in general; and Section 6 concludes the paper, presenting recommendations for future works.

2. Theoretical Foundations

2.1. Irrigation Water Quality Metrics

It is important to emphasize that drip irrigation is just one of the several available irrigation methods. It is a proven efficient method of saving water in agriculture. However, emitter clogging is a significant problem in drip irrigation systems. This problem leads to reduced application uniformity, loss of control over the applied depths, and failures in applying chemicals diluted in irrigation water.

This sensitivity to clogging is mainly affected by the dripper's characteristics and the water's quality, which are related to physical, chemical, and biological aspects [7,13,14]. Several authors have studied this problem, such as Abou-Shady et al. [15], Baeza and Contreras [16], Coelho et al. [17], and Lv et al. [18].

The problem of the emitter clogging has become the main obstacle restricting the application and promotion of drip irrigation technology, being entirely linked to the water quality for irrigation. Drip irrigation has low flow rates and extremely small passages (emitters) for water. These passages are easily clogged with organic and mineral particles from the irrigation water, chemical precipitates, and biological growth that develop within the system. For example, biofilm development is due to the mutual influence between bacterial mucilage and organic or inorganic particles. This clogging adversely affects the

performance of drip irrigation systems, resulting in less flow control and affecting the system distribution efficiency [19,20].

The main physicochemical and bacteriological parameters that provide initial insights into water characteristics, including temperature (T), pH (hydrogen potential), conductivity (EC), suspended solids (SS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), chloride, carbonate and bicarbonate, sulfate, nitrogen compounds, total aerobic mesophilic flora (TAMF), thermotolerant coliforms (TtC), total coliforms (TC), and fecal streptococci (FS) [3]. Ofori et al. [21] stated that *E. coli* and thermotolerant coliforms in water make it unsafe to irrigate vegetables and fruits.

Storlie (1995) [22] compiled the main causes of emitter clogging and the degree of restriction to water use in irrigation (Table 1).

Table 1. Degree of clogging considering different materials.

Potential Problem	Units	Degree of Restriction on Use		
		None	Slight to Moderate	Sever
Suspended solids	$mg.L^{-1}$	Less than 50	50 to 100	More than 100
pH	$mg.L^{-1}$	Less than 7	7 to 7.5	More than 7.5
Dissolved solids	$mg.L^{-1}$	Less than 500	500 to 2000	More than 2000
Manganese	$mg.L^{-1}$	Less than 0.1	0.1 to 1.5	More than 1.5
Iron	$mg.L^{-1}$	Less than 0.1	0.1 to 1.5	More than 1.5
Hardness as CaCO3	$mg.L^{-1}$	Less than 150	150 to 300	More than 300
Bacterial population	mL	10,000	10,000 to 50,000	More than 50,000

Source: [22].

The water quality index (WQI), created in 1970 by the National Sanitation Foundation (NSF) in the United States, is a formulation that enables the estimation of the overall quality of a water body based on significant parameters. The WQI evaluates raw water quality for use in public supply after treatment.

The WQI is an example of a simplified approach to assessing overall water quality by condensing abundant information into a single, typically dimensionless, value. The sodium absorption ratio (SAR) is a parameter employed to assess salinization risk due to the presence of NaCl salt resulting from irrigation [3].

The parameters used in the WQI are dissolved oxygen (DO), thermotolerant coliforms (TCs), hydrogen potential (pH), biochemical oxygen demand (oxygen consumed in 5 days at a temperature of 20 °C, BOD5,20), water temperature (Tw), total nitrogen (N-total), total phosphorus (P-total), turbidity (Tb), and total residue (Res-total). In many locations in Brazil, data on various metrics are missing [23–25].

However, different parameters are used to evaluate and monitor irrigation water quality due to the differences in its context. The parameters most commonly used for irrigation water quality assessment worldwide are electrical conductivity (EC), iron concentration (Fe), pH, and total hardness [26–31], as well as the presence and concentration of microorganisms able to cause emitter clogging.

EC can be defined as the numerical expression of the water's ability to conduct electrical current, which is related to and indicates the amount of dissolved salts in it. In general, levels greater than 100 $\mu S.cm^{-1}$ indicate impacted environments. High values may indicate water's corrosive characteristics. It is considered the most used parameter for evaluating salinity levels and concentrations of soluble salts in water for irrigation use [26,27].

Iron is one of the most abundant metals in Earth's crust. It is found in natural fresh waters at 0.5 to 50 $mg.L^{-1}$ levels. In waters containing ferrous and manganous salts, oxidation by iron bacteria (or by exposure to air) may cause rust-colored deposits on the walls of tanks, pipes, and channels and the carry-over of deposits into the water [32].

Iron is mainly present in groundwater due to its dissolution by carbon dioxide in the water. In surface waters, iron levels increase in rainy seasons due to soil transport and the occurrence of erosion processes [31]. In the State of São Paulo, a limit of 15 mg.L^{-1} was established for the concentration of soluble iron in sewage effluents discharge collection pipelines, followed by treatment [23].

pH is essential because it influences chemical balances occurring naturally or in unitary water treatment processes. The indirect effect is also significant and can, under certain pH conditions, contribute to the precipitation of chemical elements, exerting effects on nutrient solubilities [27,28].

Lastly, hardness encompasses four components: calcium bicarbonate [Ca(HCO3)2], magnesium bicarbonate [Ca(HCO3)2], calcium sulfate (CaSO4), and magnesium sulfate (MgSO4) [28]. The risk of using clogging emitters in drip irrigation is reduced when the result for this parameter is less than 150 mg.L^{-1} [29,30,33].

Finding data available for more parameters in continuous series with the proper quantity and quality is challenging. Thus, it is typical to use only a few of them in studies on water quality for irrigation, such as pH, EC, iron (Fe), and water hardness. These parameters are more accessible to measure and directly affect the emitters clogging and the formation of microorganisms related to the emitters clogging, especially iron bacteria [27–29,31].

The World Health Organization [32] considers that for microbial water quality, verification could be based on the analysis of fecal indicator microorganisms, with the organism of choice being Escherichia coli or thermotolerant coliforms. In most cases, monitoring for *E. coli* or thermotolerant coliforms provides a high degree of assurance because of their large numbers in polluted waters.

2.2. Use of AI for Water Quality Evaluation

Rahu et al. [6] emphasized that monitoring water quality is a crucial task that guarantees the safety and usability of water resources. They also considered that traditional water quality monitoring techniques take a long time, are expensive, may not be accurate, and often do not produce real-time data.

Nevertheless, it is well known that irrigation management depends on water and soil parameters. As extracting information from all the needed data to evaluate irrigation water quality is challenging, AI models and techniques could improve decision making and the management of aspects related to soil and irrigation systems.

AI encompasses state-of-the-art models used in many areas, including irrigated agriculture. Recently, many papers worldwide have reported the use of ML and deep learning algorithms in calculating or forecasting the quality of surface water and groundwater for irrigation [7,34–39].

The literature highlights the difficulties in obtaining enough data and suggests using AI to forecast the WQI. According to Nguyen et al. [38], AI models can process large amounts of data and make predictions with high precision, handling nonlinear relationships between water quality parameters, correcting missing and multidimensional data, and improving predictions as new data become available.

Singh et al. [40] evaluated the seasonal groundwater suitability for irrigation purposes using indexical approaches, statistical computing, graphical plotting, and ML algorithms. They observed that, in the context of groundwater quality prediction, multiple linear regression (MLR) and artificial neural network (ANN) models stood out, with the ANN consistently outperforming MLR. Rahu et al. [6] noted that several research studies have been conducted to investigate the use of ML algorithms in monitoring agricultural water quality. They concluded that deep learning algorithms outperform conventional regression models in terms of accuracy.

Several classifiers, including the support vector machine (SVM), random forest (RF), logistic regression (LR), decision tree (DT), CATBoost, XGBoost, and multilayer perceptron (MLP), were evaluated for classifying water quality data using the water quality index

(WQI). The study demonstrated that CATBoost achieved the highest accuracy at 94.51% compared to the other classifiers [41]. Predicting the WQI requires prior knowledge or classes, and most of them are built using expert elicitation to define the weight of each parameter in the index. Dritsas and Trigka [42] applied different classifiers in an unbalanced dataset and corroborated that, after using the Synthetic Minority Oversampling Technique (SMOTE), the classification algorithms had improved their performance.

3. Methodology Proposed

Different sources were considered to develop the proposed methodology, such as the works by Storey et al. [43] and Nafsin and Li [44]. The first work presented international studies on leading water utilities, research organizations, and technology providers worldwide involved in developing and deploying online monitoring technology to detect contaminants in water. The second work analyzed anomalous water quality events in the Milwaukee River. They provided the essential requisites that the methodology should fulfill. We also considered the main stages of the data lifecycle [45,46] and ML workflows [45]. The proposed method contains nine components, which are illustrated in Figure 1.

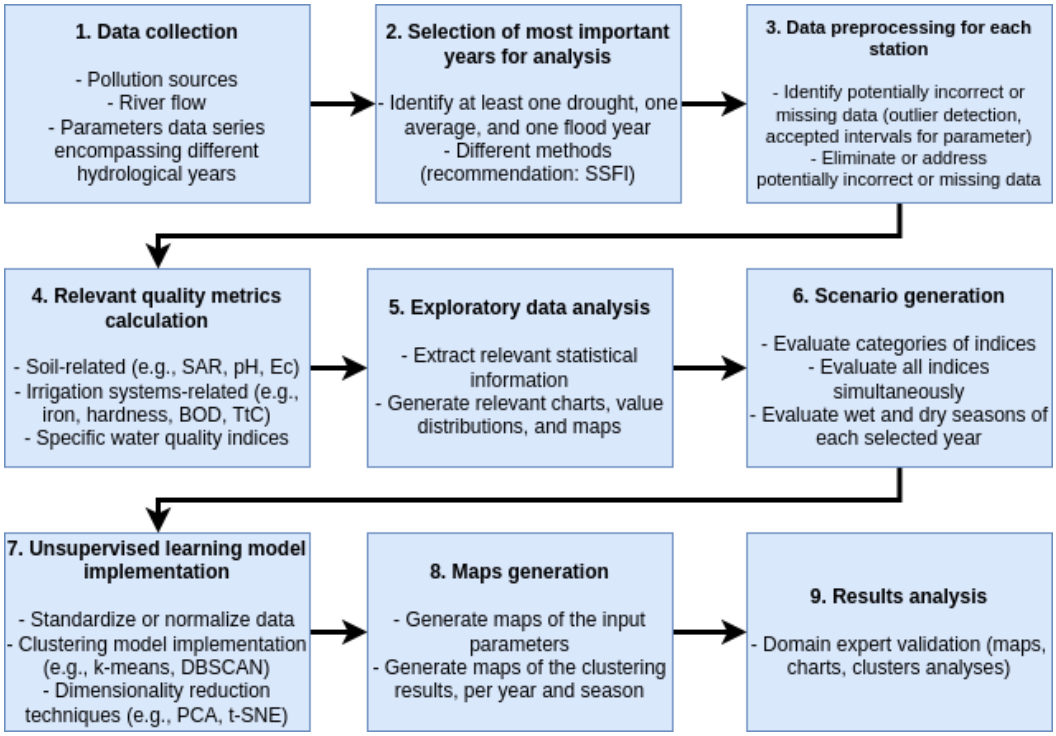

Figure 1. Proposed method's main components.

The nine components are as follows:

1. **Data collection:**This step involves collecting data for all relevant parameters and indices from all stations in the basin that encompass the regions that will be studied. For some decision-makers, the whole basin may be of interest, such as for policy making. For specific farmers, particular regions of the basin may be more critical. Although many factors could be considered, three essential ones are (i) the current presence of pollution sources (such as industries and large-size cities); (ii) the river flow; and (iii) the data series encompassing hydrological years with different characteristics (such as years of floods and of droughts);
2. **The selection of the most important years for analysis:** This step is related to selecting the hydrological years that will be analyzed and should encompass at least one flood and one drought year. We also recommend analyzing one year with average flow (which we refer to as 'average' in this paper). Several criteria and methods are used to

identify in the dataset if each year is a flood, average, or drought year. Nevertheless, we recommend using a simple and easily explainable method based on streamflow, such as the standardized streamflow index (SSFI). This method calculates the average streamflow for the whole basin for each year. Then, a criterion for identifying if the year had excess flow (indicative of a flood year) or a considerably lower flow than average (indicative of a drought year) is applied. Lastly, the years can be selected considering this classification. In the case of domain-expert selection (as was performed in this work), we recommend showing the chart of the SSFI with the triggers for flood and drought for the domain expert and then letting them select the years to be analyzed. However, the whole process can easily be automated by incorporating rules for defining flood and drought years. In the absence of previous knowledge, the years chosen may have the lowest SSFI, the highest SSFI, and the year with the SSFI closest to the average value;

3. **Data preprocessing for each station:** This aims to identify, eliminate, and address potentially incorrect or missing data. In the case of addressing missing data, different imputation methods can be used, or the sample can be discarded, depending on the specific context (with more data available, it is possible to discard data points without losing significant information). In the case of potentially incorrect data, identifying and addressing it is more challenging. We recommend identifying the accepted intervals for the parameter, considering both physical aspects (for example, pH between 0 and 14 or conductivity lower than the limit for freshwater). If data imputation is needed, several methods should be evaluated based on the value distribution for the specific parameter. The main options used are the median, average, or moving average values. If the parameter distribution is close to a normal distribution, the average is traditionally used. The median is more indicated if it differs considerably from a normal distribution.
4. **Relevant quality metric calculation:** In this step, relevant quality metrics or indices are chosen based on a literature review, legislation, or a domain expert recommendation. Several different dimensions can be considered, but the essential ones we recommend are (i) soil-related metrics, such as the sodium adsorption ratio (SAR), pH, and conductivity, which may directly influence the soil and plants; and (ii) irrigation system-related metrics, such as dissolved iron, hardness, biochemical oxygen demand, and the concentration of some microorganisms, which may cause problems such as drip clogging. As unsupervised learning models and techniques extract information directly from the data provided without prior or external knowledge, it is crucial that the dataset generated contains high-quality data. Although evaluating data quality is outside this work's scope, we refer the reader to the work by Gong et al. [47], which encompasses an in-depth review of several datasets and data quality assessment techniques and criteria.
5. **Exploratory data analysis:** After the quality metrics (also referred to as 'parameters' in this paper) are selected, they must be analyzed. This encompasses (i) extracting relevant statistical information (such as the mean, mode, median, standard deviation, and variation coefficient); (ii) generating important charts to better understand the data (such as boxplots and line charts); (iii) analyzing the value distributions for each parameter; (iv) identifying outliers; and (v) developing maps to illustrate the average values of each parameter for each season and hydrological year. This step is essential to guiding decisions such as on (i) which scenarios should be generated and evaluated; (ii) if additional data collection or processing is needed; and (iii) the potential outliers impacting the final results. Although the automation of this analysis is outside the scope of this work, it is important to emphasize that part of this evaluation can be automated, as described by Milo and Somech [48].
6. **Scenario generation:** This step aims to define and create the scenarios that will be evaluated. At least the following three aspects must be considered in different scenarios: (i) the evaluation of indices into relevant categories (such as soil-related

and irrigation system-related indices); (ii) the evaluation of all indices simultaneously; and (iii) the evaluation of the wet and dry seasons of each selected year. Additional scenarios can be generated using different unsupervised learning models and indices. Additionally, if outliers were detected during the processing or exploratory data analysis steps, it is essential to evaluate scenarios with and without outliers for each parameter that presented outlier values. This is important because sometimes the outliers are not incorrect values but extreme ones with a physical, chemical, or biological explanation. This is the case of the high concentrations of iron and biological-related parameters near populous cities. A traditional outlier detection and removal method, such as the boxplot technique, would eliminate these high values. However, they are essential to understanding water quality in the river basin in those areas. Therefore, we recommend analyzing different scenarios, such as the dataset without outliers, the dataset with all values (including outliers), and a dataset composed only of the outliers.

7. **Unsupervised learning model implementation:** In this step, an unsupervised learning model extracts valuable information from each scenario and helps generate insights for data analysis and decision making. Different methods can be used, depending on the characteristics of the data and the amount of data available. In some cases, a clustering model may be enough to extract information that improves decision making. In other cases, dimensionality reduction techniques (such as principal components analysis, PCA, or t-distributed stochastic neighbor embedding, t-SNE) can improve the results generated. However, data must be standardized or normalized before using such techniques and models, as parameters with intervals with different orders of magnitude may impact the results considerably. In general, we recommend clustering techniques to always be used in the proposed method, as one of the main objectives is to obtain and evaluate clusters of data that may bring important information related to water quality for irrigation purposes. However, in cases where there are many variables, a dimensionality reduction method is indicated to improve the exploratory data analysis and the results of the clustering model. Although there is no clear rule for what can be considered many variables, we recommend using a dimensionality reduction method if there are more than ten parameters, especially if there is the possibility that some of these parameters are partially dependent upon each other. For an in-depth evaluation of unsupervised learning methods and their applications, we refer the reader to the work by Ghahramani [11].
8. **Map generation:** This step encompasses generating maps of the parameters used as inputs for the clustering model (the indices calculated in step 4) and the clustering results used in step 7. At least one map should be generated for the clustering results for each scenario for each year or season, and one map should be generated for each parameter for each year or season. Among the options for map types that can be generated, we recommend creating one map for each parameter, separating the data into quartiles (which improves the expert validation and decision making); displaying the maps of the same parameter (or scenario) together (to make comparisons easier); displaying all scenarios together for the clustering results (to make comparisons easier); and evaluating the possibility of creating maps of differences (e.g., instead of plotting the quantile for the wet and dry seasons of a particular year as separate maps, creating a map of the quantile difference between both seasons).
9. **Result analysis:** The last step of the proposed method, which should be conducted by a domain expert with the results of the previous steps, is crucial for better decision making. In this step, the domain expert (or a group of domain experts) should compare and evaluate each parameter for each season or year (using the results of step 5), the results of the clustering of each scenario (using the results of step 7), and the maps generated (using the results of step 8). A risk analysis and temporal, spatial, and spatiotemporal analyses of each metric or group of metrics can also be conducted. This step is the most difficult to automate, as it may vary from project to project in

terms of the indices and parameters used as inputs, their distributions, the presence of outliers, the scenarios generated, and what decisions the decision-makers will make considering the results, among others.

Applying the proposed method makes it possible to address the main gaps identified in Section 1 and generate vital information to improve decision making related to water quality for irrigation purposes for different decision-makers and stakeholders. As was emphasized in the description of each step of the method, most steps can be fully automated, but some depend on domain expert analysis and validation. We believe this is not a negative aspect but a characteristic of complex problems that may impact different decisions, sustainability dimensions, and stakeholders.

It is crucial to observe that although the method is presented as a linear sequence of steps, the feedback from each step can be used to change previous steps. For example, a decision-maker in the last step may decide to generate a new scenario and incorporate or remove parameters. In this case, the whole method is applied again.

Lastly, it is important to emphasize that the code developed for the case study described in Section 4, which encompasses all the steps of the proposed method, as well as the data used, are available from an open Github repository at the following link: https://github.com/rfsilva1/data-driven-water-quality (accessed on 1 March 2024). It can be easily adapted to uses with different variables, regions, and river basins.

4. Water Quality at the PCJ Basin

This section presents the description of the case study and its main results. It is divided into the following sections: Section 4.1 describes the area and the case study itself; Section 4.2 encompasses an exploratory data analysis of each parameter; and Section 4.3 contains the main results of implementing the clustering model for each scenario and having domain experts validate the methodology.

4.1. Case Study Description

This study was conducted in the Piracicaba–Capivari–Jundiaí (PCJ) basin, one of the most critical hydrographic regions in the São Paulo state, densely populated and highly economically relevant in Brazil. Figure 2 illustrates the whole basin with its main sub-basins. The area is named after the three main rivers that form the basin, encompassing 15,304 square kilometers. The basin's annual rainfall and water flow averages are 1592 mm and 172 m^3. Few works in the water quality literature have explored this specific basin area as a whole [49].

Figure 2. PCJ river basin. Source: PCJ, 2024.

The PCJ basin fully or partially covers the territories of 76 municipalities, 71 in the state of São Paulo and 5 in the state of Minas Gerais, with an approximate length of 300 km in the east–west direction and 100 km in the north–south direction. It has an area of approximately 15,377 km², 92.45% in the state of São Paulo and 7.55% in the state of Minas Gerais, where the headwaters of the Jaguari, Camanducaia, and Atibaia rivers are located [50,51].

As for land cover, the PCJ basin has significant urban, industrial, and rural occupation playing essential roles in regional economic development. Around 7% of Brazil's Gross Domestic Product (GDP) is estimated to be produced in this area. Additionally, it encompasses a population of over 5.8 million inhabitants, with 95.1% living in urban areas (Water Resources Plan of the Hydrographic Basins of the Piracicaba, Capivari, and Jundiaí Rivers, 2020).

The Water Resources Plan for the Hydrographic Basins of the Piracicaba, Capivari, and Jundiaí Rivers [52] defined land use as follows: 25.30% encompasses pasture; 20.35% is native forest; 19.01% is sugar cane; and 12.11%, urbanized areas. Other uses are less significant in terms of the percentage of area occupied, such as other temporary and permanent crops (e.g., grain and cereal plantations, roots, vegetables, flowers, and fruits). The predominant soil types are red-yellow acrisols and red-yellow ferralsols.

Two hydrological periods were defined for this case study: (i) wet period, from 1 October of one year to 31 March of the following year; and (ii) dry period, from 1 April to 30 September of the same year. During the wet period, the release of flows to the PCJ Basins is carried out following a statement from the Department of Water and Energy (DAEE) to meet the 15-day moving average flows at the control points of the Cantareira System (SC), respecting the minimum values defined per control post.

During the dry period, an average flow of 10 m^3/s must be guaranteed, equivalent to a volume of 158.1 hm^3 to be released from the SC to the PCJ Basins [52].

The proposed methodology (Section 3) was fully implemented. The main aspects of each step are described in the following paragraphs:

1. **Data collection:** Official data were collected from the Infoáguas Online System (https://sistemainfoaguas.cetesb.sp.gov.br/), accessed on 2 February 2024. The data download encompassed the interval from 2011 to 2017, considering all stations and cities in the PCJ basin. Then, the stations located near the cities with the highest demand for irrigation were selected (as not all regions in the basin have a high demand for water for irrigation purposes). Figure 3 contains the map of the PCJ basin, illustrating the location of the stations. Although not present in the dataset, Madeira et al. [53] indicated that rivers are very high-risk quotients for pesticides and industrial chemicals, and close to 45 contaminants are present in the PCJ basin, located in an agricultural and industrial area.
2. **The selection of the most important years for analysis:** The SSFI was used to select one hydrological year for each type: higher streamflow (2011–2012, which we called 'flood'); average streamflow (2012–2013, which we called 'average'); and lower streamflow (2014–2015, which we called 'drought'). The domain experts then validated these choices. Figure 4 illustrates the SSFI calculated for the whole dataset, emphasizing the selected years.
3. **Data preprocessing for each station:** As the data were already available after an initial preprocessing, no missing data were identified. First, we aggregated the data monthly. Additionally, considering the accepted intervals for the most important indices available on the dataset, no incorrect data were detected.
4. **Relevant quality metrics calculation:** After consulting the domain experts and evaluating the data available for each quality metric, we decided to consider four relevant metrics [23,26,30,33,54] divided into two groups: (i) soil-related metrics: pH and electrical conductivity EC of water; and (ii) irrigation system-related metrics: total iron Fe, hardness, biochemical oxygen demand (BOD), and the concentration of thermotolerant coliforms (TtC). Other important metrics were lacking for most of the dataset or could not be calculated (such as the case for SAR).

5. **Exploratory data analysis:** In this step, three analyses were conducted for all parameters for each year and season: an (i) analysis of statistical information, considering the mean, median, standard deviation, minimum, and maximum values; (ii) analysis of distribution using a kernel density estimate (KDE) plot; and (iii) analysis of potential outliers using a boxplot. Additionally, maps were generated for all parameters for each year and season, separating the values into four quartiles.
6. **Scenario generation:** Three scenarios (S) were generated for each dataset (flood, average, drought) related to different metrics used as inputs for the clustering of each year. Table 2 contains the scenarios evaluated, considering their inputs and datasets.
7. **Unsupervised learning model implementation:** The k-means method is the most used clustering technique, spanning over 50 years of applications [55,56]. Therefore, it was used in the case study explored in this paper. According to Jain [55] and Steinley [56], the k-means technique has three main steps: (i) creating points to use as cluster centers in an n-dimensional space (the number of dimensions depends on the number of features on the dataset); (ii) associating all points in the dataset with the closest cluster centers (considering a specified distance metric); and (iii) recalculating the cluster centers, considering the new associations. Steps (ii) and (iii) are repeated until a stop criterion is met. This clustering method was implemented in the three scenarios. The most important hyperparameter for defining for the k-means method is the number of clusters or k. To define this hyperparameter for each scenario, three traditional methods were used: the elbow method, the dendrogram, and the silhouette score.
8. **Map generation:** Maps were generated for all the inputs for each season, as well as for the results of the clustering implementation for each scenario.
9. **Result analysis:** Two domain experts from the hydrology and irrigation domains evaluated the results generated in Steps 5, 7, and 8 while also evaluating the usefulness of the proposed methodology in relation to traditional analyses.

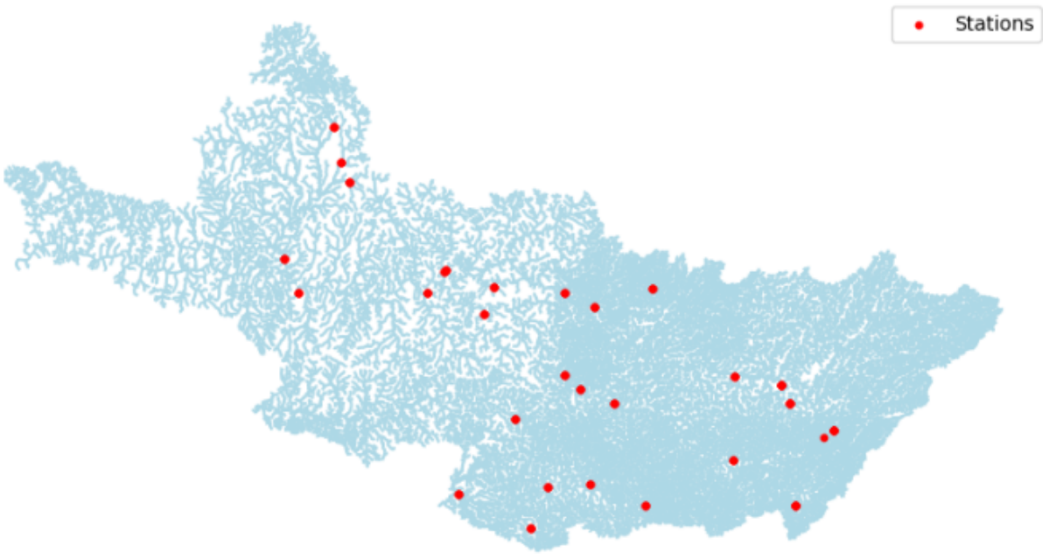

Figure 3. Data collection stations analyzed in the case study. The blue lines represent the drainage network and the red dots represent the stations

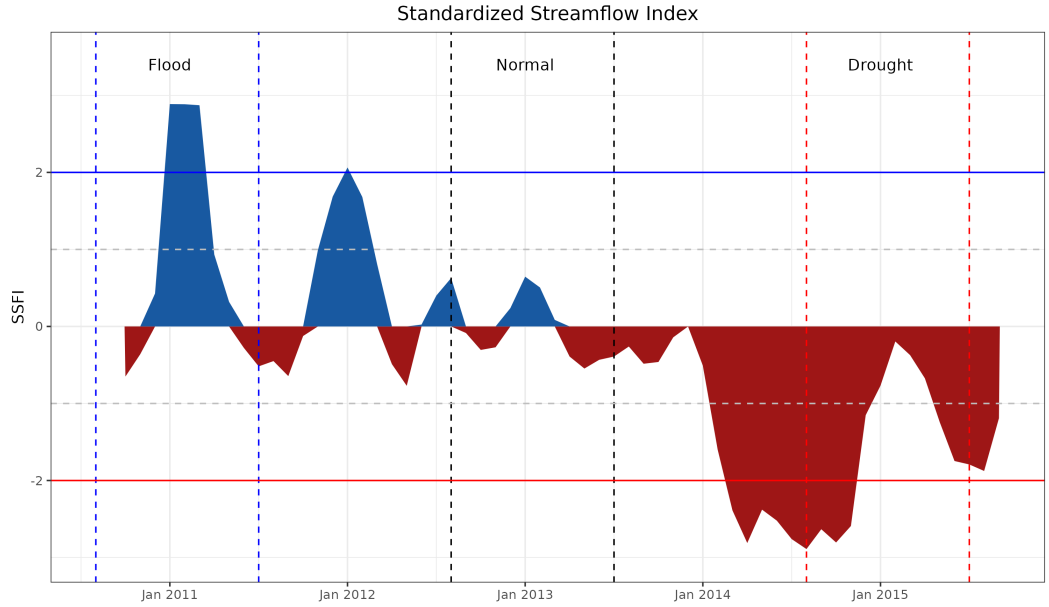

Figure 4. Standardized Streamflow Index (SSFI) calculated for the whole dataset. The blue color illustrates values of SSFI higher than 0, while the red color is related to values of SSFI lower than 0

Table 2. Case study scenarios.

Scenario	Dataset	Input Data
S11	Flood (2011–2012)	Soil-related metrics
S12	Flood (2011–2012)	Irrigation system-related metrics
S13	Flood (2011–2012)	All metrics
S21	Average (2012–2013)	Soil-related metrics
S22	Average (2012–2013)	Irrigation system-related metrics
S23	Average (2012–2013)	All metrics
S31	Drought (2014–2015)	Soil-related metrics
S32	Drought (2014–2015)	Irrigation system-related metrics
S33	Drought (2014–2015)	All metrics

Legend: soil-related metrics EC, pH; irrigation system-related metrics: Fe, hardness, BOD, TtC.

The implementation was realized using R on RStudio and Python on a Google Collaboratory CPU (https://colab.research.google.com/). The R libraries used were dplyr (https://dplyr.tidyverse.org/) [57] and lubridate (https://lubridate.tidyverse.org/) [58]. The Python libraries used were NumPy (https://numpy.org/) [59], Pandas (https://pandas.pydata.org/) [60], Matplotlib (https://matplotlib.org/) [61], Seaborn (https://seaborn.pydata.org/) [62], Scikit-Learn (https://scikit-learn.org/) [63], SciPy (https://scipy.org/) [64], GeoPandas (https://geopandas.org/) [65], and Shapely (https://shapely.readthedocs.io/). All URLs were accessed on 15 February 2024.

4.2. Exploratory Data Analysis

Table 3 shows the exploratory analysis of the different metrics considered in the case study. It is important to observe that, except for pH, all metrics present high coefficient of variation (CV) values. This amplitude occurred in all the years analyzed. The range of EC and pH values increased from the flood year to the drought year, indicating the importance of rainfall in the dilution of salts and the consequent reduction in the values of these parameters.

For Fe and hardness, the trend was the opposite, with a reduction in CV from the flood to the drought year. This indicates that the increases in rainfall and, consequently, in surface runoff sediments increase the concentrations of Fe, Ca, and Mg in the analyzed points. Both EC and pH remained within the unrestricted limits for irrigation use

(EC $\leq$ 700 µS/cm; 6 $\leq$ pH $\leq$ 8.5) [26], except for the maximum EC values during the drought year and the maximum pH value in the average year.

Although the present study cannot confirm the sources of the elements found in the water, these results suggest that further detailed analysis and monitoring are necessary to identify their origins (such as surface runoff or the disposal of organic chemical residues).

Table 4 presents the exploratory analysis of the biological parameters considered in the case study as indicators of clogging risks for irrigation systems. The CV values indicate a high data variation along the river basins, probably due to the land use (urban, industrial, agricultural), as well as the level of river streamflow. Laaraj et al. [3] observed a high spatial variation in parameters in the quality of surface water in Morocco compared to the standards established by the World Health Organization. Urban and industrial discharges had more influence over the water quality in the upper part of the watershed. The water quality improved downstream due to dilution by the streamflow of other tributaries.

Table 3. Statistical indices for the datasets for each hydrological year.

Dataset	EC ($\mu S.cm^{-1}$)	pH (U. pH)	Iron Total ($mg.L^{-1}$)	Hardness ($mg\ CaCO3.L^{-1}$)
Flood	Mean: 92.90 Std: 51.80 CV: 55.76% Min: 41.00 Max: 285.00	Mean: 6.94 Std: 0.22 CV: 3.17% Min: 6.30 Max: 7.60	Mean: 3.92 Std: 6.14 CV: 156.63% Min: 0.30 Max: 38.00	Mean: 26.49 Std: 18.04 CV: 68.10% Min: 3.48 Max: 97.00
Average	Mean: 109.55 Std: 83.63 CV: 76.34% Min: 37.30 Max: 553.00	Mean: 6.94 Std: 0.37 CV: 5.33% Min: 6.30 Max: 9.10	Mean: 2.71 Std: 3.12 CV: 115.13% Min: 0.30 Max: 19.00	Mean: 24.13 Std: 10.87 CV: 45.05% Min: 10.00 Max: 70.00
Drought	Mean: 147.43 Std: 141.21 CV: 95.78% Min: 37.00 Max: 874.00	Mean: 7.03 Std: 0.39 CV: 5.55% Min: 6.10 Max: 8.60	Mean: 1.18 Std: 0.83 CV: 70.34% Min: 0.24 Max: 6.00	Mean: 26.06 Std: 15.46 CV: 59.32% Min: 9.88 Max: 81.00

Legend: Std: standard deviation; CV: coefficient of variation; Min: minimum value; Max: maximum value.

Table 4. Statistical indices for the datasets for the biological parameters for each hydrological year.

Dataset	BOD $mg.L^{-1}$	TtC CFU
Flood	Mean: 3.46 Std: 2.08 CV: 60.16 Min: 2.00 Max: 10.00	Mean: 14602.11 Std: 29,306.59 CV: 200.70 Min: 11.67 Max: 143,333.33
Average	Mean: 3.82 Std: 1.79 CV: 46.90 Min: 2.00 Max: 8.00	Mean: 10,796.72 Std: 28,243.45 CV: 261.59 Min: 16.67 Max: 200,000.00
Drought	Mean: 4.01 Std: 2.70 CV: 67.45 Min: 2.00 Max: 14.50	Mean: 12,552.80 Std: 31770.51 CV: 253.09 Min: 1.67 Max: 250,000.00

Legend: Std: standard deviation; CV: coefficient of variation; Min: minimum value; Max: maximum value.

The analysis of the parameters in Table 5 indicates a wide range of values for most parameters, except for pH, which shows less variation (max CV = 6.71%). When looking at the data for the three years (flood, average, and drought) and their seasons (S1—wet; S2—dry), it becomes clear that a reduction in flow results in a more significant variation (with a higher CV) in the EC. The pH showed variations in both directions. Fe showed a tendency to reduce its amplitude (with a lower CV%), both from the wettest (flood) to the driest (drought) years and from the wet to the dry seasons, indicating that surface runoff is the most influential factor in the concentration of this element in the water.

According to de ALMEIDA [66], if iron concentration in water exceeds 0.1 mg.L^{-1}, precipitation issues may arise, particularly in alkaline pH. The author highlights that concentrations equal to or greater than 0.4 mg.L^{-1} can worsen the problem of iron precipitation, and the presence of iron bacteria increases the risk of clogging due to iron precipitation in the water. Nakayama and Bucks [54] demonstrated that if the iron (Fe) concentration in water exceeds 1.5 mg.L^{-1}, it can severely restrict water usage in localized irrigation. In the present case study, most measured Fe concentrations exceeded the limit, demanding an iron removal treatment method with aeration cascade before irrigating. An alkaline pH contributes to the formation of suspended solids, which easily recombine with other blockages on the emitter, increasing the clogging severity [19].

There was a downward trend in water hardness from the flood to the other two years, but the lowest values occurred in the average year. Different factors can influence this parameter, such as the extreme decrease in the water flow during the drought year and the possibility of unintentional spillage or the intentional, improper disposal of chemical and organic waste in the flood year. Given the conditions under which the present study was carried out, it is not possible to state the true causes, indicating the need for new and detailed studies to verify such hypotheses.

Table 5. Statistical indices for the datasets for each season and hydrological year.

Dataset	EC S1/S2 (µS.cm^{-1})	pH S1/S2 (U. pH)	Iron Total S1/S2 (mg.L^{-1})	Hardness S1/S2 (mg CaCO3.L^{-1})
Flood	Mean: 90.59/95.17 Std: 45.61/57.61 CV: 50.35%/60.53% Min: 42.50/41.00 Max: 270.00/285.00	Mean: 6.93/6.95 Std: 0.24/0.20 CV: 3.46%/2.88% Min: 6.30/6.50 Max: 7.40/7.60	Mean: 5.46/2.41 Std: 7.86/3.22 CV: 143.96%/133.61% Min: 0.90/0.30 Max: 38.00/16.00	Mean: 31.00/22.06 Std: 23.27/8.97 CV: 75.07%/40.66% Min: 3.48/8.00 Max: 97.00/47.00
Average	Mean: 108.38/110.73 Std: 78.89/88.94 CV: 72.79%/80.32% Min: 37.30/42.70 Max: 424.00/553.00	Mean: 6.88/7.00 Std: 0.20/0.47 CV: 2.91%/6.71% Min: 6.30/6.50 Max: 7.30/9.10	Mean: 3.57/1.85 Std: 3.90/1.72 CV: 109.24%/92.97% Min: 0.60/0.30 Max: 19.00/12.00	Mean: 23.96/24.29 Std: 10.81/11.05 CV: 45.12%/45.49% Min: 10.00/12.00 Max: 61.00/70.00
Drought	Mean: 142.16/154.44 Std: 150.86/128.25 CV: 106.12%/83.04% Min: 37.00/39.00 Max: 874.00/661.00	Mean: 7.04/7.01 Std: 0.43/0.35 CV: 6.11%/4.99% Min: 6.20/6.10 Max: 8.32/8.60	Mean: 1.27/1.07 Std: 0.95/0.62 CV: 74.80%/57.94% Min: 0.30/0.24 Max: 6.00/3.00	Mean: 24.87/27.66 Std: 16.21/14.39 CV: 65.18%/52.02% Min: 9.88/10.30 Max: 81.00/73.00

Legend: Std: standard deviation; CV: coefficient of variation; Min: minimum value; Max: maximum value.

Parameters BOD and TtC (Table 6) show high values but different patterns for CV. In BOD, CV was higher in the drought year and lower in the average year, and the mean values increased from the flood to the drought year. In TtC, CV was higher in the average year and lower in the flood year, and the mean values were higher in the flood year and lower in the average year. The mean values of BOD indicate that water availability directly influences the dilution of organic materials in the rivers. Otherwise, the mean values of TtC are inconclusive, as TtC was higher in the flood year and lower in the average year, suggesting many possibilities that were not investigated in the present study. Such possibilities include

streamflow contact with materials with high concentrations of microorganisms in the flood year and the increased discharge of some substances during flood years.

Sreekala et al. [67] found a similar pattern in a study in Central Kerala, India. According to the authors, the seasonal changes in the total and thermotolerant coliform concentrations could be due to rainfall, overland flow, nutrient load, and temperature change. Boithias et al. [68] analyzed twelve flash-flood events sampled from 2011 to 2015 at the outlet of a tropical montane headwater catchment in Northern Lao, using *E. coli* as a fecal indicator bacteria, aiming to quantify the contributions of both surface runoff and sub-surface flow to the in-stream concentration of fecal coliforms during flood events. The authors found that the highest concentrations of *E. coli* occur in flood events driven by surface runoff, concluding that surface runoff and soil erosion are the primary drivers of in-stream fecal coliform contamination.

Table 6. Statistical indices for the datasets for the biological parameters for each season and hydrological year.

Dataset	BOD S1/S2 mg.L^{-1}	TtC S1/S2 CFU
Flood	Mean: 3.72/3.19 Std: 3.70 /1.86 CV: 99.50/58.84 Min: 2.00/2.00 Max: 19.00/9.00	Mean: 14,107.87/15,086.04 Std: 22,781.35/34,775.29 CV: 161.48/230.51 Min: 86.67/11.67 Max: 110,000.00/143,333.33
Average	Mean: 4.17/3.48 Std: 2.59/2.24 CV: 62.10/64.40 Min: 2.00/2.00 Max: 11.00/11.00	Mean: 11014.76/10578.68 Std: 31,233.93/25,232.95 CV: 283.56/238.53 Min: 50.00/16.67 Max: 200,000.00/160,000.00
Drought	Mean: 4.33/3.39 Std: 5.95/2.07 CV: 137.32/61.05 Min: 2.00/2.00 Max: 50.00/10.00	Mean: 11,755.69/13,615.62 Std: 27,652.73/36,799.34 CV: 235.23/270.27 Min: 1.67/1.67 Max: 166,666.67/250,000.00

Legend: Std: standard deviation; CV: coefficient of variation; Min: minimum value; Max: maximum value.

Figure 5 illustrates the distribution of iron concentration and TtC in the data analyzed during the flood, average, and drought years. One notes that the highest iron concentrations (>35 mg.L^{-1-1}) rarely occurred, and most data showed concentrations from 0 to 5 mg.L^{-1}. Even so, iron concentrations above 0.3 mg.L^{-1} are enough to cause clogging events in drip irrigation systems. Clogging problems can be aggravated by the presence of microorganisms known as iron bacteria, which, in high concentrations of iron, form mucilage on the walls of pipes and drippers. High iron concentrations in the water favor the growth of bacteria and, hence, result in clogging problems in drip irrigation systems.

Most data found TtC concentrations to be equal to or less than 20 × 103 UFC 100 mL^{-1}. According to Storlie et al. (1995) [22], these levels define the degree of restriction on the use for irrigation as "Slight to Severe" (Table 1). The problem seems higher in the flood year, which presents the highest average population, even if the maximum population occurred during the drought year (250 × 103 UFC 100 mL^{-1}).

One important tool requested by the experts to improve the quality of the analysis of each metric considering multiple seasons and years was a map representing the values of the parameters for each station. However, as potential outliers exist in the data (as observed in the previous analysis), we opted to develop maps for all parameters for each season and year, aggregating the data into quantiles. This allowed us to reduce the impact of potential outliers in the analysis and obtain a less challenging visualization of variations between seasons for each parameter.

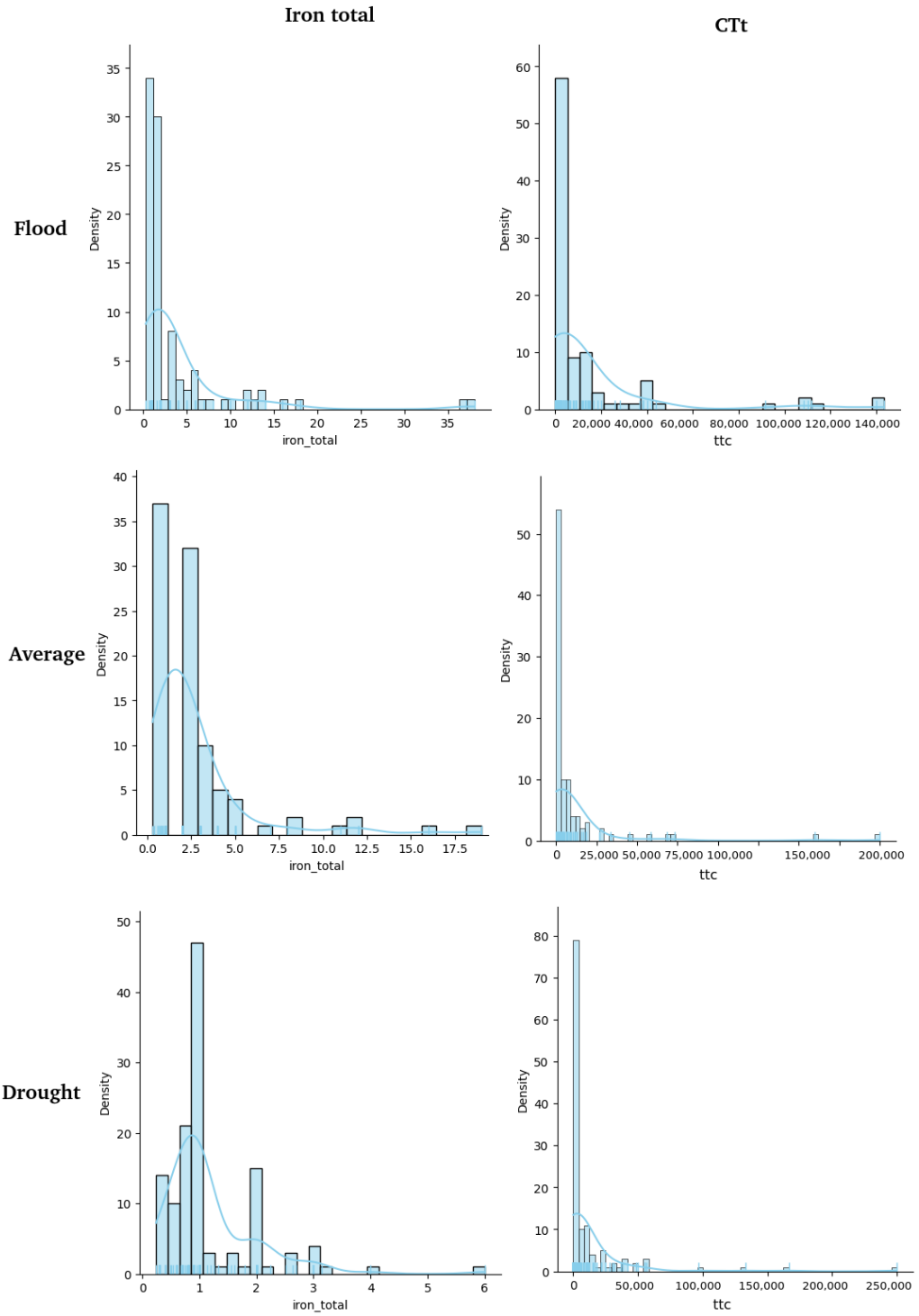

Figure 5. KDE distributions of the iron total and TtC parameters for the different datasets.

Figure 6 provides an example of such an analysis for the total iron metric. One can easily note that the changes in iron concentration between the seasons, but they are not so clear between the years analyzed, especially in the drought year. Otherwise, one can detect the changes in the flood year, denoting the influence of the water flow level on iron concentration.

Figure 6 also emphasizes the slight variation between iron contents in 2014 (drought year) because the rivers kept their flow low throughout the year. Therefore, the rainy season differed very little from the dry period.

This highlights the importance of performing additional analyses using more than one parameter to understand the data. All the maps and parameter distribution charts (including boxplot and KDE distribution) are available if requested.

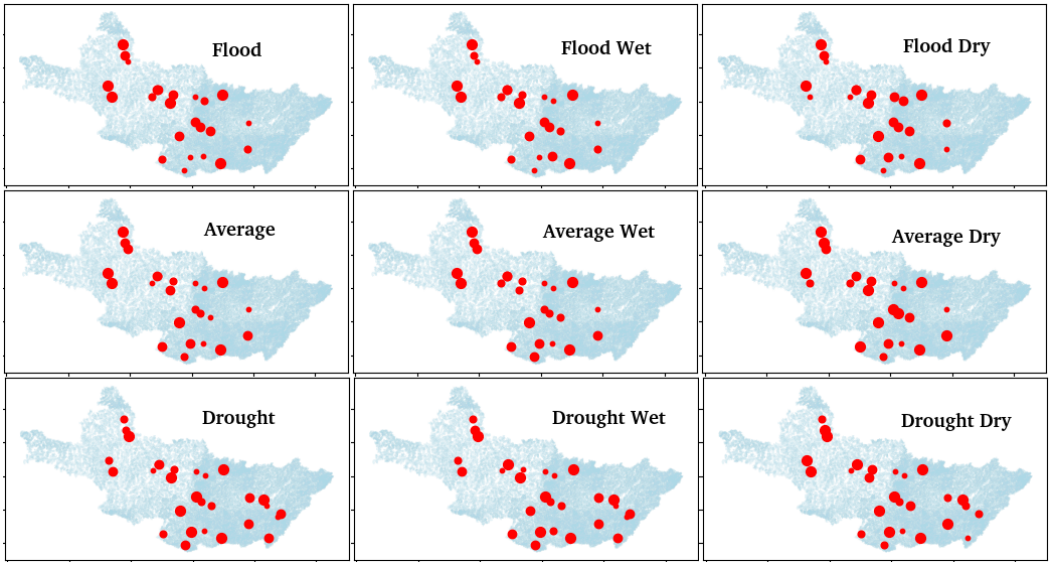

Figure 6. Iron concentrations in the different seasons and years, grouped by quantiles. The quantile of iron concentration is represented in red, and the drainage network is represented in light blue. Legend: the bigger the circle size, the higher the concentration.

We obtained similar results in analyzing the thermotolerant coliforms metric (Figure 7). It is easier to detect changes in TtC between seasons in the same year than between years. Nevertheless, such detection between different years was easier with TtC (Figure 7) than with iron (Figure 6).

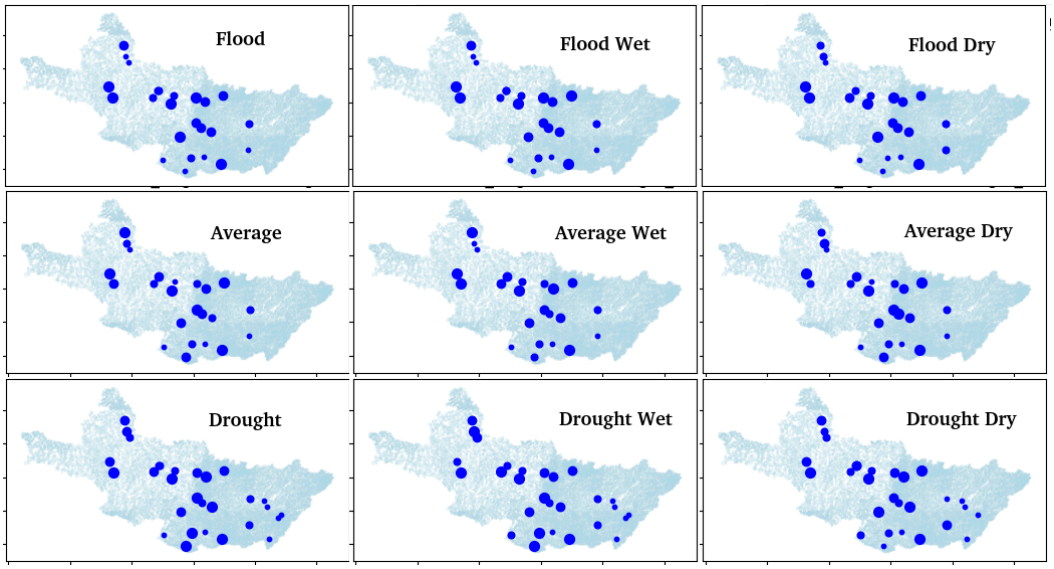

Figure 7. Original TtC in the different seasons and years, grouped by quantiles. The quantile of TtC is represented in dark blue, and the drainage network is represented in light blue. Legend: the bigger the circle size, the higher the concentration.

4.3. Result Analysis and Domain Expert Validation

The result analysis and validation (the last step of the proposed method, described in Section 3) were conducted by domain experts analyzing the maps generated (both of inputs and clusters, as shown in Section 4.2), the distributions of inputs, and statistical index tables

generated automatically by applying the framework. The clusters were generated using one of the most widely used clustering models (k-means clustering), which considered different techniques to determine the number of clusters in the data. k-means is a purely data-driven method that does not depend on experts' inputs. Then, experts evaluated the results of the method and the exploratory data analysis conducted by applying our proposed method (described in Section 3).

Answering RQ1, which is related to the identification of the most critical components for developing a data-driven methodology to analyze water quality for irrigation considering the soil and irrigation system dimensions, we observed that nine components are fundamental, considering the evaluation of domain experts: data collection; the selection of the most important years for analysis; data processing; quality metric calculation; exploratory data analysis; scenario generation; unsupervised learning model implementation; map generation; and result analysis.

These components allow for a more automated analysis, providing essential insights for decision making. Additionally, several components can be further explored to generate different scenarios, extract information using different artificial intelligence models and techniques, and generate meaningful insights for improving decisions from different agents. They also encompass the most critical stages of the data lifecycle, considering aspects from data collection to decision making. Lastly, these components can be easily implemented in the traditional ML workflow.

Research question 2, related to the variation in the water quality for the three hydrological years considered, requires observing that a decrease in iron concentration is expected during the dry season due to less soil transport. This effect can be observed for the average and drought year maps in Figure 6. However, it cannot be observed for the flood year, when a smaller flow variation is expected between the dry and wet seasons.

The smaller flow variation in the flood year can also explain the lower number of clusters in that year when compared to the average and dry years (Figure 8). This effect could also be observed if the sizes of the sub-basins were evaluated. It can also be observed that, as the study region has a sub-humid climate, the EC and pH parameters vary relatively little and can be grouped into a smaller number of clusters in the flood year (Figure 8).

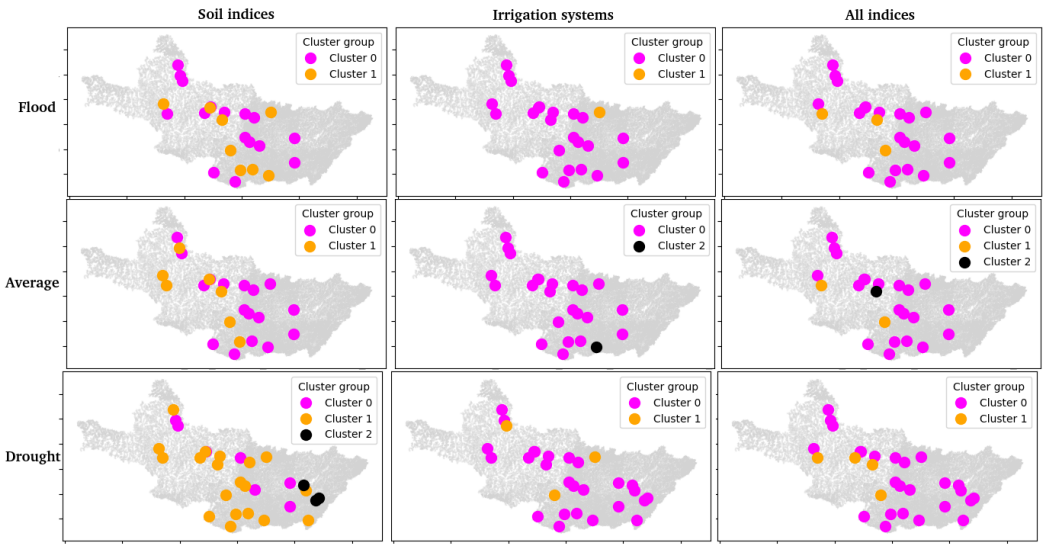

Figure 8. Clustering results for all datasets and scenarios.

5. Discussion

Based on the case study results, it is possible to observe that the clustering process can result in interesting insights if it is necessary to select a smaller number of points to be sampled at some point during a basin analysis. In this context, the proposed methodology

can generate important maps and help to cluster the data considering different sub-basins, seasons, and hydrological years.

The number of clusters found (between two and three in the different scenarios) was considered by the domain experts as enough to describe the most important aspects of water quality for irrigation purposes for the area analyzed in the case study. The method separated the traditionally more polluted areas, such as the ones in the western and southern regions of the basin, mainly due to discharges from densely populated areas (as illustrated in Figure 7). An interesting aspect is that this was accomplished without prior expert knowledge, only using the available data on the dataset.

It is important to emphasize that the focus of the proposed method in this case study was on data-driven knowledge extraction for a better understanding of the problem and providing important information for decision making and analysis, not to predict water quality itself (although this could be a follow-up study, using supervised learning methods such as artificial neural networks, support vector machines, and random forests, among other methods).

Therefore, the proposed method is an alternative to traditional analyses, which usually depend on experts' inputs and objective definitions for specific river basins. The only exception was the case of TtC, which is highly relevant in basins with a high concentration of urban areas (and that proved itself a very relevant variable for clustering, as it had locations with a considerably high number of colony-forming units, influencing the clustering process significantly). We adopted a three-class classification before using the data as an input for the clustering method. However, raw values were used for the other variables, so the framework is not dependent on specific variables. Its main limitation is the need to use tabular data, an important aspect of future research.

Additionally, it is important to observe that it can help to better study and evaluate the potential impacts of climate change on the different sub-basins and regions. Climate change can increase the threat to soil and irrigation systems. Global warming will likely increase river water temperature, causing direct impacts on biogeochemical processes in freshwater [69]. Compared to increasing streamflows, temperature was the primary driver of the increased release of manganese from river beds [70].

Singh et al. [40] evaluated groundwater quality for use in irrigation based on electrical conductivity (EC), residual sodium carbonate (RSC), the sodium adsorption ratio (SAR), sodium concentration (Na%), and water quality index (WQI). They concluded that seasonal dynamics influence groundwater quality for irrigation, demonstrating notable changes in the concentrations of cations and anions. For surface waters, the variation must be more significant. Anthropogenic disturbance and land use are most likely crucial in determining spatial water quality patterns in rivers [9].

According to the report presented by the Water Resources Plan for the Piracicaba, Capivari, and Jundiaí River Basins [52], in 2014, the levels of the Sistema Cantareira reached historic lows. This supports the results observed in the case study in the present work. From 2013 onward, there has been a considerably dry rainfall regime in the region, where water input has been the lowest ever recorded in the system, which for the first time, had an entire year with water inputs lower than outputs.

In April 2013, the water volume began to fall. On 7 July 2014, removal through the pumping of the first portion of the dead volume was authorized. In November 2014, authorization was given to remove a second quota of the dead volume. On 4 February 2015, the system reached its historic low. The dead volume continued to be used until December 2015, when rains in the region restored volumes to values above the dead volume. In March 2016, available volumes continued to increase, reaching operational normality [52].

According to the report presented by the Water Resources Plan for the Piracicaba, Capivari, and Jundiaí River Basins [52], impermeable areas reduce the infiltration and natural recharge of aquifers. Urban surfaces become polluted by the deposits of airborne contamination and solid wastes. This contamination is transported to rivers by rain, contaminating the river, especially at the beginning of the rain periods. Due to the increases

in runoff speed and unprotected surfaces, there are substantial increases in sediments in urban rivers and the erosion of banks, mainly altering the iron content in the water.

From a practical aspect, the proposed method can provide important information to improve decision making for any variable or quality index available and any area or river basin, as long as the data used are in tabular format. As cited before, the need to provide data in a tabular format is one of the main limitations of the proposed framework, as data in grid format or tensors with more than two dimensions cannot be directly used with the code available.

However, from a theoretical perspective, it is still possible to use clustering and dimensionality reduction techniques that work with spatiotemporal data, which have this characteristic. For an overview of this problem, potential solutions, and different models and techniques that could be used, we refer the reader to the works by Ansari et al. [71] and Shi and Pun-Cheng [72].

Another limitation of the present work is that only one area was evaluated in the study case (even though this area encompasses several important rivers and may provide critical information for different stakeholders and decision-makers). It would be interesting to apply the method to basins with different characteristics (e.g., basin area, climate, indices and parameters used, and decisions that will be made using the information provided by the method, among others) in different countries. This would provide important feedback to improve the method, which is an effort already being made

6. Conclusions and Future Works

Evaluating and monitoring water quality is essential for using irrigation efficiently, improving productivity, and addressing sustainability concerns. However, there is a gap in the literature for automated methods for implementing a pipeline for water quality evaluation and monitoring that considers all stages of the data lifecycle and the traditional ML workflow. Therefore, the main objectives of this work were to propose such a data-driven method, consider different water quality indices, and conduct a case study implementing it with real data.

This work had two main research questions. The first was identifying the main components necessary to propose such a data-driven method. We identified and implemented nine components, from data collection considering multiple sources to providing maps, tables, and important charts to improve decision-making. The second question relates to the case study results. We concluded that in the flood year, possibly due to less variation in weather, the parameters that describe the salinization risk could be grouped into fewer clusters. The proposed methodology can also enhance domain experts' analysis and improve decision making.

Additionally, the proposed methodology can be used with different parameters and quality indices and does not depend directly on prior domain experts' knowledge to provide important information to improve decision making. It can also be applied to different areas and river basins. Several options for automation and variations in its components were explored throughout this work.

The following are recommendations for future works: exploring additional data processing techniques, such as PCA and t-SNE; exploring additional unsupervised learning models, such as the density-based spatial clustering of applications with noise (DBSCAN) and self-organizing maps (SOMs); accounting for the flow measured at the time of collection and the size of the sub-basins; evaluating additional water quality metrics, such as SAR, aluminum, and manganese; developing a web app based on the code implemented for the case study to allow for better and faster water quality monitoring; and evaluating the use of spatiotemporal clustering models for improving decision making.

Author Contributions: Conceptualization, R.F.d.S., M.R.B., F.E.C., T.G.M., F.C.M., P.A.A.M., E.M.M., A.C.B.D. and A.M.S.; methodology, R.F.d.S. and M.R.B.; software, R.F.d.S. and M.R.B.; writing—original draft preparation, R.F.d.S., M.R.B., F.C.M., P.A.A.M. and S.N.D.; writing—review and editing,

R.F.d.S., M.R.B. and P.A.A.M.; supervision, E.M.M., A.C.B.D. and A.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by: the Sao Paulo Research Foundation (FAPESP grant #2019/07665-4), IBM Corporation, and the National Council for Scientific and Technological Development (CNPq). We also acknowledge the support from the Coordenacão de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, finance code 001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study can be downloaded from the following repository: https://github.com/rfsilva1/data-driven-water-quality (accessed on 1 March 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zuffo, A.C.; Duarte, S.N.; Jacomazzi, M.A.; Cucio, M.S.; Galbetti, M.V. The Cantareira System, the Largest South American Water Supply System: Management History, Water Crisis, and Learning. *Hydrology* **2023**, *10*, 132. [CrossRef]
2. Lopes, T.R.; Folegatti, M.V.; Duarte, S.N.; Moster, C.; Zolin, C.A.; Oliveira, R.K.; Moura, L.B. Economic value of environmental services regulating flow and maintaining water quality in the Piracicaba River basin, Brazil. *J. Water Resour. Plan. Manag.* **2023**, *149*, 05023008. [CrossRef]
3. Laaraj, M.; Benaabidate, L.; Mesnage, V.; Lahmidi, I. Assessment and modeling of surface water quality for drinking and irrigation purposes using water quality indices and GIS techniques in the Inaouene watershed, Morocco. *Model. Earth Syst. Environ.* **2024**, *10*, 2349–2374. [CrossRef]
4. Wu, B.; Tian, F.; Zhang, M.; Piao, S.; Zeng, H.; Zhu, W.; Liu, J.; Elnashar, A.; Lu, Y. Quantifying global agricultural water appropriation with data derived from earth observations. *J. Clean. Prod.* **2022**, *358*, 131891. [CrossRef]
5. Soares, S.R.A.; Fontenelle, T.H.; Ferreira, D.A.C.; Gonçalves, M.V.C.; Dourado Neto, D.; Barretto, A.G.d.O.P.; Fendrich, A.N.; Safanelli, J.L.; Araujo, M.A.d.; Coutinho, P.A.Q.; et al. *Atlas Irrigação: Uso da água na Agricultura Irrigada*; ANA:Brasília, Brazil 2021.
6. Rahu, M.A.; Shaikh, M.M.; Karim, S.; Chandio, A.F.; Dahri, S.A.; Soomro, S.A.; Ali, S.M. An IoT and machine learning solutions for monitoring agricultural water quality: A robust framework. *Mehran Univ. Res. J. Eng. Technol.* **2024**, *43*, 192–205. [CrossRef]
7. Egbueri, J.C.; Mgbenu, C.N.; Digwo, D.C.; Nnyigide, C.S. A multi-criteria water quality evaluation for human consumption, irrigation and industrial purposes in Umunya area, southeastern Nigeria. *Int. J. Environ. Anal. Chem.* **2023**, *103*, 3351–3375. [CrossRef]
8. Pereira, M.A.; Marques, R.C. Sustainable water and sanitation for all: Are we there yet? *Water Res.* **2021**, *207*, 117765. [CrossRef] [PubMed]
9. Wu, Z.; Lai, X.; Li, K. Water quality assessment of rivers in Lake Chaohu Basin (China) using water quality index. *Ecol. Indic.* **2021**, *121*, 107021. [CrossRef]
10. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *An Introduction to Statistical Learning: With Applications in Python*; Springer Nature: Berlin/Heidelberg, Germany, 2023.
11. Ghahramani, Z. Unsupervised learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, 2–14 February 2003, Tübingen, Germany, 4–16 August 2003, Revised Lectures*; Springer: Berlin/Heidelberg, Germany. 2004; pp. 72–112.
12. Aliashrafi, A.; Zhang, Y.; Groenewegen, H.; Peleato, N.M. A review of data-driven modelling in drinking water treatment. *Rev. Environ. Sci. Bio/Technol.* **2021**, *20* , 985–1009. [CrossRef]
13. Muniz, G.L.; Duarte, F.V.; Rakocevic, M. Assessment and optimization of carbonated hard water softening with moringa oleifera seeds. *Desalin. Water Treat* **2020**, *173*, 156–165. [CrossRef]
14. Muniz, G.L.; Camargo, A.P.; Signorelli, F.; Bertran, C.A.; Pereira, D.J.; Frizzone, J.A. Influence of suspended solid particles on calcium carbonate fouling in dripper labyrinths. *Agric. Water Manag.* **2022**, *273*, 107890. [CrossRef]
15. Abou-Shady, A.; Siddique, M.S.; Yu, W. A Critical Review of Innovations and Perspectives for Providing Adequate Water for Sustainable Irrigation. *Water* **2023**, *15*, 3023. [CrossRef]
16. Baeza, R.; Contreras, J.I. Evaluation of thirty-eight models of drippers using reclaimed water: Effect on distribution uniformity and emitter clogging. *Water* **2020**, *12*, 1463. [CrossRef]
17. Coelho, R.D.; de Almeida, A.N.; de Oliveira Costa, J.; de Sousa Pereira, D.J. Mobile drip irrigation (MDI): Clogging of high flow emitters caused by dragging of driplines on the ground and by solid particles in the irrigation water. *Agric. Water Manag.* **2022**, *263*, 107454. [CrossRef]
18. Lv, C.; Niu, W.; Du, Y.; Sun, J.; Dong, A.; Wu, M.; Mu, F.; Zhu, J.; Siddique, K.H. A meta-analysis of labyrinth channel emitter clogging characteristics under Yellow River water drip tape irrigation. *Agric. Water Manag.* **2024**, *291*, 108634. [CrossRef]

19. Li, R.; Han, Q.; Dong, C.; Nan, X.; Li, H.; Sun, H.; Li, H.; Li, P.; Hu, Y. Effect and Mechanism of Micro-Nano Aeration Treatment on a Drip Irrigation Emitter Based on Groundwater. *Agriculture* **2023**, *13*, 2059. [CrossRef]
20. Perboni, A. Sensibilidade de Gotejadores à Obstrução por Partículas de Areia. *Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo*. 2016. Available online: https://irriga.fca.unesp.br/index.php/irriga/article/view/2162 (accessed on 15 February 2024).
21. Ofori, S.; Abebrese, D. K.; Ruzickova, I.; Wanner, J. Reuse of Treated Wastewater for Crop Irrigation: Water Suitability, Fertilization Potential, and Impact on Selected Soil Physicochemical Properties. *Water* **2024**, *16*, 484. [CrossRef]
22. Storlie, C. Treating Drip Irrigation System with Chlorine. *Ruthgers Cooperative Extension Services Fact Sheet FS795*. 1995. Available online: https://njaes.rutgers.edu/FS795/ (accessed on 30 January 2024).
23. CETESB. Apêndice D: Índices de Qualidade das Águas. 2020. Available online: https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2020/09/Apendice-D-Indices-de-Qualidade-das-Aguas.pdf (accessed on 30 January 2024).
24. Gradilla-Hernández, M.S.; de Anda, J.; Garcia-Gonzalez, A.; Montes, C.Y.; Barrios-Piña, H.; Ruiz-Palomino, P.; Díaz-Vázquez, D. Assessment of the water quality of a subtropical lake using the NSF-WQI and a newly proposed ecosystem specific water quality index. *Environ. Monit. Assess.* **2020**, *192*, 296. [CrossRef] [PubMed]
25. Abdul Maulud, K.N.; Fitri, A.; Wan Mohtar, W.H.M.; Wan Mohd Jaafar, W.S.; Zuhairi, N.Z.; Kamarudin, M.K.A. A study of spatial and water quality index during dry and rainy seasons at Kelantan River Basin, Peninsular Malaysia. *Arab. J. Geosci.* **2021**, *14*, 1–19. [CrossRef]
26. Ayers, R.S.; Westcot, D.W. *Water Quality for Agriculture*; Food and agriculture organization of the United Nations Rome, FAO: Rome, Italy 1985; Volume 29.
27. Aliyu, T.; Balogun, O.; Namani, C.; Olatinwo, L.; Aliyu, A. Assessment of the presence of metals and quality of water used for irrigation in Kwara State, Nigeria. *Pollution* **2017**, *3*, 461–470.
28. Aminiyan, M.M.; Aitkenhead-Peterson, J.; Aminiyan, F.M. Evaluation of multiple water quality indices for drinking and irrigation purposes for the Karoon river, Iran. *Environ. Geochem. Health* **2018**, *40*, 2707–2728. [CrossRef] [PubMed]
29. Malakar, A.; Snow, D.D.; Ray, C. Irrigation water quality—A contemporary perspective. *Water* **2019**, *11*, 1482. [CrossRef]
30. Muniz, G.L.; Oliveira, A.L.G.; Benedito, M.G.; Cano, N.D.; Camargo, A.P.d.; Silva, A.J.d. Risk Evaluation of Chemical Clogging of Irrigation Emitters via Geostatistics and Multivariate Analysis in the Northern Region of Minas Gerais, Brazil. *Water* **2023**, *15*, 790. [CrossRef]
31. Singh, V.K.; Bikundia, D.S.; Sarswat, A.; Mohan, D. Groundwater quality assessment in the village of Lutfullapur Nawada, Loni, District Ghaziabad, Uttar Pradesh, India. *Environ. Monit. Assess.* **2012**, *184*, 4473–4488. [CrossRef] [PubMed]
32. World Health Organization. WHO Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First Addendum. 2017. Available online: https://iris.who.int/bitstream/handle/10665/254637/9789241549950-eng.pdf?sequence=1 (accessed on 20 May 2024).
33. de Almeida, O. *Qualidade da água de Irrigação*; Embrapa Mandioca e Fruticultura: Cruz das Almas, Brazil, 2010.
34. Aminu, I.I. A novel approach to predict water quality index using machine learning models: A review of the methods employed and future possibilities. *Glob. J. Eng. Technol. Adv.* **2022**, *13*, 026–037. [CrossRef]
35. Babbar, R.; Babbar, S. Predicting river water quality index using data mining techniques. *Environ. Earth Sci.* **2017**, *76*, 1–15. [CrossRef]
36. Giri, S. Water quality prospective in Twenty First Century: Status of water quality in major river basins, contemporary strategies and impediments: A review. *Environ. Pollut.* **2021**, *271*, 116332. [CrossRef] [PubMed]
37. Mokhtar, A.; Elbeltagi, A.; Gyasi-Agyei, Y.; Al-Ansari, N.; Abdel-Fattah, M.K. Prediction of irrigation water quality indices based on machine learning and regression models. *Appl. Water Sci.* **2022**, *12*, 76. [CrossRef]
38. Nguyen, D.P.; Ha, H.D.; Trinh, N.T.; Nguyen, M.T. Application of artificial intelligence for forecasting surface quality index of irrigation systems in the Red River Delta, Vietnam. *Environ. Syst. Res.* **2023**, *12*, 24. [CrossRef]
39. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R.; Kumar, S. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **2021**, *276*, 130265. [CrossRef] [PubMed]
40. Singh, G.; Singh, J.; Wani, O.A.; Egbueri, J.C.; Agbasi, J.C. Assessment of groundwater suitability for sustainable irrigation: A comprehensive study using indexical, statistical, and machine learning approaches. *Groundw. Sustain. Dev.* **2024**, *24*, 101059. [CrossRef]
41. Nasir, N.; Kansal, A.; Alshaltone, O.; Barneih, F.; Sameer, M.; Shanableh, A.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. *J. Water Process Eng.* **2022**, *48*, 102920. [CrossRef]
42. Dritsas, E.; Trigka, M. Efficient data-driven machine learning models for water quality prediction. *Computation* **2023**, *11*, 16. [CrossRef]
43. Storey, M.V.; Van der Gaag, B.; Burns, B.P. Advances in on-line drinking water quality monitoring and early warning systems. *Water Res.* **2011**, *45*, 741–747. [CrossRef] [PubMed]
44. Nafsin, N.; Li, J. Using CANARY event detection software for water quality analysis in the Milwaukee River. *J. Hydro-Environ. Res.* **2021**, *38*, 117–128. [CrossRef]
45. De Silva, D.; Alahakoon, D. An artificial intelligence life cycle: From conception to production. *Patterns* **2022**, *3*, 100489. [CrossRef] [PubMed]

46. Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data lifecycle challenges in production machine learning: A survey. *ACM SIGMOD Rec.* **2018**, *47*, 17–28. [CrossRef]
47. Gong, Y.; Liu, G.; Xue, Y.; Li, R.; Meng, L. A survey on dataset quality in machine learning. *Inf. Softw. Technol.* **2023**, *162*, 107268. [CrossRef]
48. Milo, T.; Somech, A. Automating exploratory data analysis via machine learning: An overview. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, Portland, OR, USA, 14–19 June 2020; pp. 2617–2622.
49. Silva, A.A.F.d.; Esteves, K.E. Ecological and biological patterns of stream fish studies from the Piracicaba-Capivari-Jundiaí Basin (PCJ Basin, SP) assessed through a systematic review. *Biota Neotrop.* **2023**, *23*, e20221440. [CrossRef]
50. ANA. Agência Nacional de Águas e Saneamento Básico: PCJ. 2024 Available online: https://www.gov.br/ana/pt-br/assuntos/gestao-das-aguas/planos-de-recursos-hidricos/planos-de-recursos-hidricos-de-bacias-hidrograficas/planos-de-bacias-hidrograficas-interfederativas/pcj (accessed on 20 February 2024).
51. PCJ. Agência das Bacias do PCJ. 2024. Available online: https://agencia.baciaspcj.org.br/bacias-pcj/localizacao/ (accessed on 20 February 2024).
52. das Bacias PCJ, C.P. Plano de Recursos Hídricos das Bacias Hidrográficas dos Rios Piracicaba, Capivari e Jundiaí, 2020 a 2035: Relatório Final./Executado por Consórcio Profill-Rhama e Organizado porComitês PCJ/Agência das Bacias PCJ. 2024. Available online: https://drive.google.com/file/d/1Vom4DKOTzTnvrIKOmEJtZlPMzScAcOOe/view (accessed on 20 February 2024).
53. Madeira, C.L.; Acayaba, R.D.; Santos, V.S.; Villa, J.E.; Jacinto-Hernández, C.; Azevedo, J.A.T.; Elias, V.O.; Montagner, C.C. Uncovering the impact of agricultural activities and urbanization on rivers from the Piracicaba, Capivari, and Jundiaí basin in São Paulo, Brazil: A survey of pesticides, hormones, pharmaceuticals, industrial chemicals, and PFAS. *Chemosphere* **2023**, *341*, 139954. [CrossRef]
54. Nakayama, F.; Bucks, D. *Trickles Irrigation for Crop Production*; US Department of Agriculture, Agricultural Research Service, US Water Conservation Laboratory: Phoenix, AZ, USA, 1986; p. 383.
55. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
56. Steinley, D. K-means clustering: A half-century synthesis. *Br. J. Math. Stat. Psychol.* **2006**, *59*, 1–34. [CrossRef] [PubMed]
57. Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D. dplyr: A Grammar of Data Manipulation. 2023. Available online: https://dplyr.tidyverse.org (accessed on 15 February 2024).
58. Grolemund, G.; Wickham, H. Dates and Times Made Easy with lubridate. *J. Stat. Softw.* **2011**, *40*, 1–25. [CrossRef]
59. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. . [CrossRef] [PubMed]
60. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Stéfan van der Walt, S., Millman, J., Eds.; 2010; pp. 56–61. . [CrossRef]
61. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. . [CrossRef]
62. Waskom, M.L. seaborn: statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. . [CrossRef]
63. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
64. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
65. Jordahl, K.; den Bossche, J.V.; Fleischmann, M.; Wasserman, J.; McBride, J.; Gerard, J.; Tratner, J.; Perry, M.; Badaracco, A.G.; Farmer, C.; et al. geopandas/geopandas: v0.8.1. *Zenodo* **2020** . . [CrossRef]
66. de ALMEIDA, O. *Entupimento de Emissores em Irrigação Localizada*; Embrapa Mandioca e Fruticultura Tropical: Cruz das Almas, Brazil, 2009.
67. Sreekala, M.; Sareen, S.J.; Rajathi, S. Influence of Geo-environmental and Chemical Factors on Thermotolerant Coliforms and *E. coli* in the Groundwater of Central Kerala. *J. Geol. Soc. India* **2018**, *91*, 621–626. [CrossRef]
68. Boithias, L.; Ribolzi, O.; Lacombe, G.; Thammahacksa, C.; Silvera, N.; Latsachack, K.; Soulileuth, B.; Viguier, M.; Auda, Y.; Robert, E.; et al. Quantifying the effect of overland flow on Escherichia coli pulses during floods: Use of a tracer-based approach in an erosion-prone tropical catchment. *J. Hydrol.* **2021**, *594*, 125935. [CrossRef]
69. Liu, S.; Xie, Z.; Liu, B.; Wang, Y.; Gao, J.; Zeng, Y.; Xie, J.; Xie, Z.; Jia, B.; Qin, P.; et al. Global river water warming due to climate change and anthropogenic heat emission. *Glob. Planet. Change* **2020**, *193*, 103289. [CrossRef]
70. Paufler, S.; Grischek, T.; Benso, M.R.; Seidel, N.; Fischer, T. The impact of river discharge and water temperature on manganese release from the riverbed during riverbank filtration: A case study from Dresden, Germany. *Water* **2018**, *10*, 1476. [CrossRef]
71. Ansari, M.Y.; Ahmad, A.; Khan, S.S.; Bhushan, G.; Mainuddin. Spatiotemporal clustering: A review. *Artif. Intell. Rev.* **2020**, *53*, 2381–2423. [CrossRef]
72. Shi, Z.; Pun-Cheng, L.S. Spatiotemporal data clustering: A survey of methods. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 112. [CrossRef]