Discover Water

Research

Assessing the impact of missing data on water quality index estimation: a machine learning approach

David Sierra-Porta[1]

Received: 14 July 2023 / Accepted: 6 March 2024
Published online: 14 March 2024

Abstract

Despite the regulations and controls implemented worldwide by governments and institutions to ensure the availability and quality of water resources, many water sources remain susceptible to contamination. This contamination poses significant risks to human health and can lead to substantial economic losses. One of the challenges in this context is the presence of missing or incomplete data, which can arise from various factors such as the methodology used or the expertise of personnel involved in sample collection and analysis. The existence of such data gaps hampers the accurate analysis that can be conducted. To address this issue and estimate a water quality index from the available samples, it is crucial to handle missing information appropriately to avoid biased calculations. This study focuses on the application of machine learning methods for imputing missing data in water samples. Furthermore, it quantifies the performance of different models based on the distribution of the obtained data. By applying 10 distinct methods to a sample of water quality data, the most effective approaches, namely Bayesian Ridge, Gradient Boosting, Ridge, Support Vector Machine, and Theil-Sen regressors, were identified. The selection of these models was based on the evaluation of two estimation error metrics: average percent bias (PBIAS) and Kling-Gupta Efficiency statistic (KGEss). The respective metric values for the aforementioned methods are as follows: $\langle \text{PBIAS} \rangle_{0.5} = 14.665, 19.555, 14.300, 15.380, 15.920$ and $\langle \text{KGEss} \rangle_{0.5} = 0.670, 0.585, 0.655, 0.620, 0.595$. The results obtained from these models have been utilized to establish unbiased relationships among physical, chemical, and biological parameters based on the information retrieved through the applied imputation methods.

Keywords Water quality · Imputation methods · Machine learning · Data mining · Process improvement

1 Introduction

Water, as one of the most vital mineral resources crucial for human existence, plays a significant role in our development, survival, and commercial activities. Despite its importance, water resources are often underutilized and not fully optimized. Considering the abundance of water on the planet (refer to Table 1), approximately 95–97% of it exists in the form of saltwater in seas and oceans [1–3]. Consequently, this salty water cannot be directly consumed, utilized for agricultural purposes, or employed in most industrial processes. The remaining 3–5% of water is fresh, but a significant portion is locked in ice at the polar regions and glaciers.

✉ David Sierra-Porta, dporta@utb.edu.co | [1]Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar., Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias 130010, Bolívar, Colombia.

Table 1 Distribution of water resource across of globe

Resource	Volume (km^3 × 10^{-6})	Total water (%)	Fresh water (%)
Atmospheric water	0,0129	0,001	0,01
Glaciers	24,064	1,72	68,7
Ground Ice	0,3	0,021	0,86
Rivers	0,00212	0,0002	0,006
Lakes	0,1764	0,013	0,26
Swamps	0,01147	0,0008	0,03
Soil Moisture	0,0165	0,0012	0,05
Aquifers	10,53	0,75	30,1
Lithosphere	23,4	1,68	–
Oceans	1338	95,81	–

The table appear in [1]. Water on Earth is primarily distributed among three categories based on its salinity. Approximately 97.5% of the water on the planet is saline, predominantly found in oceans and seas. The remaining 2.5% is freshwater, with about 68.7% trapped in the form of ice in glaciers and polar ice caps. The remaining 0.3% constitutes the freshwater available in rivers, lakes, and underground aquifers

Moreover, approximately 67–69% of the water derived from glaciers is fresh, while around 30% originates from aquifers and other sources. Consequently, only a minuscule proportion of 0.003% of the total water mass on Earth is readily accessible and suitable for human use.

Each country bears the responsibility for managing sanitation and purification processes to ensure the highest possible quality of water, particularly for human consumption. In light of these circumstances, governments and institutions dedicated to the conservation and monitoring of water quality must establish methodologies to ensure the correct application of minimum sanitary standards, thereby guaranteeing its quality.

To achieve this goal, these institutions commonly collect water samples from various sources, such as outfalls, rivers, lakes, dams, and other ecosystems. Subsequently, these samples are transported to laboratories where they undergo thorough analysis and are subjected to diverse measurement and qualification processes.

Water quality is influenced both naturally by the characteristics of the river basin and artificially by human activities within that basin [4–7]. The concept of quality is closely related to the specific uses for which the water is intended. Thus, distinctions are made between the quality required for domestic use, which differs from the quality necessary for human consumption, irrigation, or ecosystem support. Furthermore, water quality encompasses the conditions that water must meet to maintain a balanced ecosystem. Water pollution occurs when foreign substances, often resulting from human activities, are introduced. Pollutants can take various forms, including chemicals, organic matter, soil particles, and plastics [8, 9]. Additionally, any discharge occurring in secondary basins is likely to find its way into rivers, eventually reaching dams and potentially altering the water content used for human consumption.

Real-time water quality monitoring data are invaluable for conducting innovative studies that address dynamic temporal variations, such as water quality prediction, assessment, and environmental management [10].

In the evaluation of water quality, particularly in monitoring and environmental management systems, a common challenge, especially in real-time and automated processes, is the prevalent issue of missing data [11–13]. This lack of data often arises from equipment failures, network coverage limitations, or data corruption, constituting a frequently encountered problem in studies and systems where not all data or parameters are reported on time, measurement instrumentation and equipment may be unavailable, or data tabulation issues may arise [14]. The loss of data introduces randomization dilution, unknown biases, and compromises the statistical power of studies and analyses, posing a serious challenge to the reliability of results. Missing data may pertain to the effectiveness of treatments, adverse effects, or prognosis. However, research papers or reports often do not explicitly address or specify the extent of missing data in their studies, and many computer programs assume that data are complete, further exacerbating the problem.

The issue of missing data is pervasive across decision-making processes and information management in general. Analyzing data in the absence of relevant information can lead to erroneous conclusions or, at best, conclusions unsupported by robust indicators [15]. Moreover, many advanced analysis techniques require complete data, as the algorithms rely on real values (or strings) for all instances and variables in the dataset [16]. Therefore, it is crucial to develop effective approaches for handling missing data.

Missing data refers to the absence of values that would be meaningful or useful for result analysis. There are various types of missing data, and multiple reasons can account for their occurrence, significantly impacting how missing data should be addressed during result analysis. One key consideration is determining whether the missingness is random, affecting all individuals equally, or whether it is due to specific reasons that could introduce biases, potentially invalidating the results.

In the existing literature, various approaches have been proposed to address the issue of missing data in water quality studies. For instance, artificial neural networks have been utilized, as demonstrated in the work of Tabari et al. [17]. The study employed a combination of artificial neural networks, specifically, MultiLayer Perceptron (MLP) [18, 19] and Radial Basis Function (RBF) [20, 21] algorithms to reconstruct missing river water quality data. The research found that the MLP and RBF algorithms were effective in estimating important parameters, such as hardness, based on other water quality parameters.

While the use of MLP and RBF networks in water quality studies offers advantages, it comes with certain disadvantages. RBF networks require a larger number of neurons for training, especially with a high number of training vectors, leading to increased computational effort and memory storage requirements. Additionally, selecting centers for RBF networks can be challenging, unlike MLP networks where this process does not occur. The complex task of selecting centers in RBF networks can impact the network's performance and training process. Training an MLP network can be time-consuming when the output dimension is high, whereas RBF networks are less influenced by the output dimension. However, RBF networks lack extrapolation capability, returning a result of 0 far away from the centers of the RBF layer, limiting their use for extrapolation. In the context of missing components, MLP networks are less affected if a weight or neuron is missing, while RBF networks can experience a strong local error for a lesion and a weak but global error, impacting the network's output. These limitations may restrict the flexibility and adaptability of RBF or MLP networks in certain applications, particularly in addressing missing data in water quality. While these techniques offer valuable capabilities, their complexities and limitations should be considered to ensure their effective and appropriate use in addressing missing data and enhancing decision-making in water quality management.

Another approach involves the use of Hot-Deck Methods, as showcased in the study by Srebotnjak et al. [22]. This research highlights how these imputation methods, when applied to freshwater quality parameters, can enhance decision-making by incorporating geographic coverage information. This suggests that while the hot-deck imputation method offers a valuable approach to address missing data, it is essential to recognize and address the limitations associated with the availability and quality of geographical data to ensure the reliability and accuracy of the imputed values and the resulting water quality index. Without robust geographical data, the effectiveness and accuracy of the hot-deck imputation method may be compromised, leading to potential biases in the imputed values and the overall water quality index. Therefore, the availability and quality of geographical data pose a significant challenge in the successful application of the hot-deck imputation method for water quality studies.

Additionally, machine learning techniques have been employed in various imputation methods for predicting water quality parameters with a high percentage of missing values, as shown by Rodriguez et al. [23]. However, the study faced certain disadvantages and difficulties, particularly related to the high percentage of missing data and the high temporal and spatial variability in the water-quality datasets. The substantial proportion of missing data posed a significant challenge, as traditional imputation methods may not be well-suited to handle such a large volume of missing values. Additionally, the high temporal and spatial variability in the datasets may have introduced complexities in the imputation process, as the imputed values needed to accurately capture the dynamic and heterogeneous nature of water quality parameters across different time periods and geographical locations. Moreover, effectively due to the large amount of missing data in the dataset used in their study, the authors needed to use supporting variables to proceed with the process of recovering missing data. In other words, recovery of missing values is justified, but it is necessary to utilize many other features and variables that are complete, indicating that much missing data requires much more additional information.

However, it is important to note that imputation methods, including those developed using machine learning tools, should not be regarded as the ultimate goal of the study. Instead, their purpose is to provide missing information to enhance other processes, such as improving the predictability of data-driven recommender systems. Many of the previous studies referenced in the preceding paragraphs primarily focus on evaluating such methods as a standalone objective. In our case, not only compare the performance of missing data imputation methods but also utilize this information to assess the percentage improvement in modeling and prediction processes before and after applying the imputation method, specifically for predicting the standard water quality index.

The forthcoming study explore techniques for recovering and inferring missing data in real datasets obtained from water quality risk systems used to determine drinking water quality indices and other everyday applications. Leveraging original and official information provided by agencies involved in monitoring aqueduct systems that supply drinking water to the population in Colombia, evaluating the effectiveness of applying imputation methods to address information gaps in generating models for determining drinking water quality.

The methods outlined above pose a significant challenge due to their intricate implementation requirements. Difficulties arise from the extensive computing time needed for computational experiments and the necessity for profound knowledge of computer language and programming. However, here it's present a different perspective in this argument. This study, adopt a much simpler approach, recognizing that operators and data analysts in Latin American cities, particularly in Colombia [24], often lack the technical expertise for the advanced analyses proposed in previous studies. Moreover, computational resources and the requisite expertise for developing machine learning models are frequently unavailable, especially in rural areas distant from major urban centers. This situation is further complicated by the time constraints faced by our analysts, who need prompt insights into the specific conditions of locations where water quality data is under scrutiny.

In contrast, our study employs techniques that are easily implementable on various information management and analysis platforms. These techniques do not demand advanced technical proficiency for implementation. Platforms such as Amazon Web Services, Google Cloud, and Microsoft offer highly configurable and user-friendly interfaces with pre-trained machine learning models. These platforms enable users, including water quality analysis centers, to access and utilize these resources with minimal training or induction.

Specifically, to address the previous challenge, the paper utilize data collected by SIVICAP (*Sistema de Información de la Vigilancia de la Calidad del Agua para Consumo Humano*) to determine the quality conditions of water for human consumption in Colombia. From the data collected over a wide time interval, it's recognize the percentage of missing data in the dataset and demonstrate how using machine learning tools can improve conditions for a more efficient estimation of the water quality index locally. This enhancement benefits decision-makers by providing a more solid basis for decision-making.

In contrast to certain prior investigations, our study distinguishes itself through its noteworthy contribution, involving the assessment of the efficacy of models employed for data imputation. This contribution extends to demonstrating how these models enhance predictions and elucidate established connections between physico-chemical and biological parameters. By establishing these relationships, our study aids in making well-informed decisions rooted in improved models.

The paper is divided into the following parts. Sections 2.1 and 2.2 present the data used and the collection sources, as well as the data mining and preprocess engineering. Sections 2.3 and 2.5 outline the methods used to recover the missing data in the dataset and the performance metric to evaluate the strategy. In Sect. 3, the results of the implementations are discussed, and finally, in Sect. 4, the conclusions and final comments are presented.

2 Data and methods

To conduct this study, a methodology is employed based on extracting data from the monitoring systems and analyzing water samples from reservoirs and dams intended for human consumption by the Colombian government. Given the substantial amount of missing data in the available variables of the physico–chemical–biological tests of the samples, established a methodology to recover these values, enabling more robust analysis and improving prognosis and decision models.

The process begins with a data mining and data engineering phase, where preprocess, arrange, clean, and organize the information. Unimportant variables are discarded, and the most suitable set of physicochemical–biological variables is established. Subsequently, selecting several machine learning algorithms to model the recovery of variables from the data. Performance metrics and model evaluations are employed. Finally, the models are used to create an imputed dataset with information on the missing variables, which is then utilized to establish water quality prediction models for human consumption. The entire process is schematically represented in Fig. 1.

2.1 Water quality index

According to Ball and Church [25], water quality indices can be classified into 10 categories grouped into 4 main groups. These categories are organized based on their specific uses. For a comprehensive list of these groups and categories, please refer to the original publication by [25].

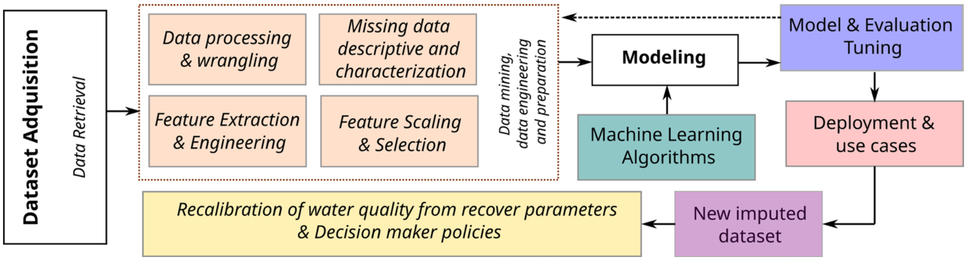
Dataset Adquisition
Data Retrieval
Data processing & wrangling
Missing data descriptive and characterization
Feature Extraction & Engineering
Feature Scaling & Selection
Data mining, data engineering and preparation
Modeling
Model & Evaluation Tuning
Machine Learning Algorithms
Deployment & use cases
Recalibration of water quality from recover parameters & Decision maker policies
New imputed dataset

Fig. 1 Methodology used in this study. This study employs a methodology involving data extraction from monitoring systems and analysis of water samples from Colombian government-designated reservoirs and dams for human consumption. Due to significant missing data, a methodology is established, involving data mining, engineering, and machine learning, to recover values in physico–chemical–biological tests, enhancing analysis and decision models for water quality prediction

In the context of human health risk assessment, one of the most commonly employed types of indices is the drinking water quality index. Each country, through its respective evaluation systems and designated institutes responsible for water resource monitoring, establishes its own definition and utilization of indices for assessing drinking water quality.

Most water quality indexes (WQIs) follow a similar calculation structure. They involve standardizing the parameters incorporated within the index based on their concentrations and subsequently assigning weights to these parameters based on their importance in determining the overall perception of water quality. The calculation of WQIs involves integrating the weighted parameters using various mathematical functions [26].

There are two primary approaches to calculating WQIs. The first approach utilizes a weighted product method, where the weights assigned to the scores of each parameter are multiplied together, considering their respective importance. The second approach employs a weighted sum method, where each score is multiplied by its corresponding weight, and the products are then summed to obtain the index value. In the case of equal weights assigned to each score, the weighted sum approach simplifies the calculation.

The index value is called the unweighted arithmetic value, if the sum of the weights is not equal, it is known as the arithmetic value of water quality [25]. Some examples for WQIs calculation equations, associated by groups according to the type of equation used are: (a) Geometric weighted indices, NFS-WQI (EU) [27, 28], Dinius (EU) [29], CETESB (Brazil) [30] in which $\mathrm{WQI}_m = \prod_{i=1}^{n} I_i^{W_i}$, where W_i is the weight or percentage assigned to the i-th parameter, I_i is the subscript of the i-th parameter; (b) Arithmetic weighted indices as for example used for UWQI (Europe) [31, 32] which $\mathrm{UWQI}_m = \sum_{i=1}^{n} W_i I_i$ where W_i is the weight or percentage assigned to the i-th parameter and I_i is the subscript of the i-th parameter; (c) Mixed weighted indices like CCME-WQI (Canada) [33–35], DWQI (EU), $\mathrm{CCME}_m = 100 - 0,5773\sqrt{F_1^2 + F_2^2 + F_3^2}$, the index incorporates three elements: Scope (F_1): percentage of parameters exceeding the standard, Frequency (F_2): percentage of individual tests of each parameter exceeding the standard, Amplitude (F_3): magnitude by which each parameter that does not meet the standard exceeds the standard; and (d) Custom truncated indices used for example in ISQA (Spain) [36], ISQA_m =T(COD+DO+SS+Cond) where T is Temperature, COD: Chemical Oxygen Demand, DO: Dissolved Oxygen, Cond: Conductivity, SS: Suspended Solids.

In Colombia, the National Institute of Health operates the Information System for Monitoring the Quality of Water for Human Consumption (SIVICAP, Sistema de Información de la Vigilancia de la Calidad del Agua para Consumo Humano), which serves as the health authority responsible for reporting water quality monitoring data. The primary indicator or index used to assess water quality for human consumption is the "Indice de Riesgo de Calidad del Agua" (IRCA), which translates to the Water Quality Risk Index in English.

The measurement methodology for IRCA is similar to other indices such as the UWQI (Water Quality Index) with some variations. The purpose of IRCA is to assess the risk associated with the non-compliance of water with physical, chemical, and microbiological standards established for human consumption. The IRCA is categorized into different ranges, each corresponding to a specific level of risk: 0–5% (No risk-Water suitable for human consumption), 5.1–14% (Low level of risk), 14.1–35% (Medium level of risk), 35.1–70% (High level of risk), and 70.1–100% (Unfeasible from a sanitary perspective).

The calculation of IRCA involves a weighted average, where risk scores are assigned to each characteristic (physical, chemical, and microbiological) based on their impact on water quality and health risks. The numerator represents the sum of scores assigned to characteristics that do not meet the quality parameters, while the denominator represents the total sum of scores for all analyzed characteristics [24].

By utilizing this weighted average calculation, the IRCA provides an assessment of the overall risk level associated with water quality, helping to prevent the occurrence of diseases related to water consumption that does not comply with established standards.

$$\mathrm{IRCA} = \frac{\sum^{\mathrm{NAC}} \mathrm{RISK}_{score}}{\sum^{\mathrm{TC}} \mathrm{RISK}_{score}} \times 100, \tag{1}$$

where NAC sum running over Non-Acceptable Characteristics and TC sum running over the complete Total of Characteristics.

The risk factors (RISK_{score}) are characterized by the norm standard and are used to code each of the measurements in the chemical and biological elements in the water and used to determine water quality according to acceptability conditions in accordance with national and international standardized regulations.

2.2 Data

Data used in this study come from the Water Quality Risk Index for Human Consumption (https://www.datos.gov.co/api/odata/v4/wppj-n4q2) for 27 municipalities in the Department of Caldas Colombia (Lat = 5°06′ N, Lon = 75°33′ O). Caldas is one of the thirty-two departments that, together with Bogotá, the Capital District, form the Republic of Colombia. Its capital is Manizales. It is located in the center of the country, in the Andean region, bordering Antioquia to the north, Boyacá to the northeast, Cundinamarca to the east, Tolima and Risaralda to the south and Risaralda to the west. With 7888 km^2 it is the fifth least extensive department and with 125 inhabitants/km^2, the sixth most densely populated.

It belongs to the coffee-growing region and the Paisa region. It was created in 1905, as a result of the reform of the political-administrative division at that time. In this department it is possible to find all the thermal floors, from the warm valleys of the Magdalena and Cauca rivers to the perpetual snows of the Nevado del Ruiz. The mountainous topography is predominant.

The data consists of 216 instances (27 municipalities for 8 years of sampling) and 131 attributes corresponding to measurements of 20 different parameters, namely: Total Alkalinity, Aluminum, Free Residual Chlorine, Chlorides, Total Coliforms, Apparent Color, Conductivity, Total Organic Carbon (TOC), Cryptosporidium, Total Hardness, Florides, E.coli, Fluorides, Giardia, Total Iron, Nitrites, Odor, ph, Taste, Sulfates, Turbidity, for average, maximum and minimum values in addition to the index calculated for these measurements (IRCA, Water Quality Risk Index for Human Consumption or *Índice de Riesgo de la Calidad del Agua para Consumo Humano* by its Spanish acronyms). The number of all samples and the number of Not-Suitable samples and their corresponding percentages are also collected.

For the purposes of this study, 56 variables have been retained, corresponding to the data for the number of samples, averages, and maximums. In addition to the averages, information about the number of samples to obtain the average of each measurement and the maximum value of that measurement for each physicochemical–biological parameter has been included. Other variables, such as minimum measurements, have been eliminated due to an almost total loss of data. Table 2 displays the percentage of missing data for 23 variables in the dataset.

From Table 2, it is evident that at least half of the considered variables contain missing data. Among these, 46% of the variables exhibit a minimum of 20% missing data, with an additional 6 variables surpassing the threshold of 30% missing data. The distribution of missing data is assumed to be random and lacks a standardized pattern. The complete set of variables with missing data, on average, lacks approximately 20% of information.

Although not employed in the methodology of this study, Fig. 2 illustrates the distribution of the Water Quality Index measure for all monitoring stations in each municipality, categorized by the year of evaluation. According to the results, all the values measured in each year indicate samples that pose a high risk for human consumption.

2.3 Imputation methods and missing data handle

The presence of missing data is a common challenge faced by researchers and decision-makers. While having a complete dataset is ideal, applying inappropriate imputation methods can create more problems than solutions. Over the past few decades, alternative procedures with better statistical properties have been developed, surpassing traditional options such as listwise deletion, pairwise deletion, mean imputation, and hot-deck imputation. Multiple imputation (MI) algorithms have been introduced and can be implemented using various commercial and freely available packages. It is important to note that the choice of imputation method should be tailored to each

Table 2 Percentage of missing values throughout the 23 data attributes in the dataset

Variable	MD (%)	Variable	MD (%)
Nitrites	88.89	Total Iron Average	20.37
Maximum Odor	47.22	Maximum Sulfates	18.52
Average Odor	47.22	Sulfates Average	18.52
Maximum Nitrites	43.06	Clorides Average	15.74
Taste	40.74	Maximum Clorides	15.74
Maximum Taste	40.74	Maximum Total Alkalinity	15.28
Maximum Aluminium	32.41	Maximum Total Hardness	15.28
Average Aluminium	32.41	Total Hardness Average	15.28
Fluorides Average	28.70	Total Alkalinity Average	15.28
Maximum Fluorides	24.54	Maximum TOC	8.33
Maximum Total Iron	20.37	TOC Average	8.33
Untreated Samples	0.46		

MD: Missing Data. Values in MD are written in percentages respect total of data

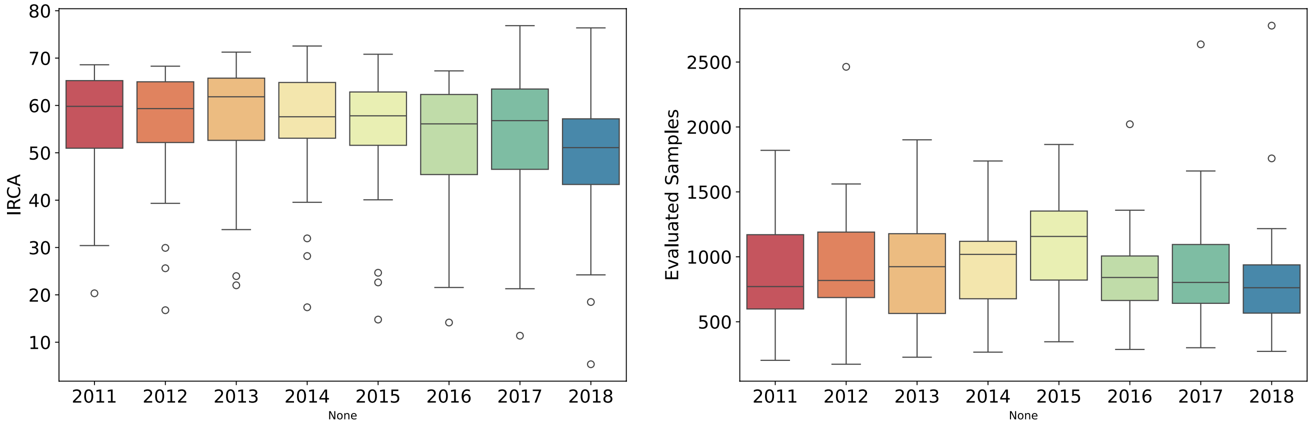

Fig. 2 BoxPlot for Water Quality Index in Colombia and Number of Samples evaluated for determination of index and desegregated by year

specific situation, considering the non-response rate and its spatial distribution, as these characteristics may vary between surveys.

In the context of water quality research, it is crucial to acknowledge and address the challenges posed by biases, missing data, and their subsequent imputation. Bias, often arising from systematic errors in data collection or analysis, can significantly impact the reliability of study outcomes. Missing data, a common occurrence in real-world datasets, introduces complexities in the analysis, potentially leading to skewed interpretations. This study recognizes the need for a comprehensive understanding of these concepts to ensure the robustness of water quality assessments.

Biases in water quality data can arise from various sources, including sampling methods, instrumentation, and environmental factors. To enhance the significance of our research, we must delve into the implications of biases on the interpretation of water quality parameters. Clear identification and mitigation of biases ensure that the results obtained are more representative of the true state of water quality, reinforcing the reliability and applicability of our findings.

Missing data, whether due to logistical constraints or other factors, presents a common hurdle in water quality studies. It is imperative to acknowledge the impact of missing values on the comprehensiveness of our analyses. This study employs advanced imputation techniques to address missing data, ensuring a more complete dataset for accurate assessments. By elucidating our approach to handling missing values, we aim to enhance transparency and foster a deeper understanding of the methodology employed.

The statistical imputation process plays a pivotal role in refining our dataset by estimating missing values. However, it is essential to communicate the intricacies of this process to maintain transparency and foster trust in the results. This study employs machine learning algorithms for imputation, specifically emphasizing Bayesian Ridge, Gradient Boosting, Ridge, Support Vector Machine, and Theil-Sen regressors. Understanding these imputation methods is crucial for interpreting the reliability of water quality predictions and the subsequent impact on decision-making processes.

2.4 Machine learning approach to missing data

This section presents the methods employed in this study to construct a model for imputing missing and non-response values related to the physical, chemical, and biological composition of water samples for human consumption. This research is part of a broader project aimed at evaluating the efficacy of Machine Learning techniques in imputing missing data for both quantitative and qualitative variables. To achieve this objective, the study conduce imputation experiments using several Machine Learning algorithms briefly described below.

In general, these methods leverage the information present in the dataset to create prediction models using a classification process that employs decision trees for stratification and discrimination of target variables based on the conditions of predictor variables. Some of the models, in addition to decision trees, utilize bootstrap methods to reduce variance and generate more decision trees, thereby increasing the information available for each classification iteration. Additionally, some models incorporate non-parametric regression into the classification task by combining the initial predictions with decision trees to enhance the predictive power on the target variables.

All the algorithms employed in this study are implemented in a Python library [37] specifically designed for these purposes. The sklearn library (https://scikit-learn.org/stable/) [38, 39], along with the modules sklearn.ensemble and sklearn.linear_model, encompasses all these algorithms.

2.4.1 RandomForestRegressor (RF)

RF is a machine learning algorithm [40, 41] that is based on the ensemble learning technique, where multiple models are combined to improve overall performance. The philosophy behind Random Forest lies in the idea that combining multiple decision trees can reduce the overfitting inherent in a single tree and improve the generalizability of the model.

The algorithm works by building a "forest" of decision trees during training. Each tree is trained on a random subsample of the original data set and a random selection of features, ensuring that the individual trees are diverse and not overly correlated with each other. During prediction, the results of all the trees are combined to produce a final prediction, either by averaging the predictions in the regression case or voting in the classification case.

RF outperforms traditional techniques based on robust statistics alone by taking advantage of the diversity and complexity of multiple decision trees. While traditional statistical methods may face limitations in the ability to capture nonlinear relationships or complex interactions between variables, Random Forest can model these relationships more effectively thanks to its ability to handle high-dimensional data and nonlinearities.

In terms of performance optimization, the most important hyperparameters in RF include the number of trees in the "forest", the maximum depth of each tree, and the minimum number of samples required to split a node. Adjusting these hyperparameters appropriately can help avoid overfitting and improve the generalizability of the model.

2.4.2 AdaBoostRegressor (AB)

AB is a machine learning algorithm that belongs to the boosting family of methods [42, 43]. The underlying idea behind AdaBoostRegressor is to build a strong predictive model from the weighted combination of multiple weaker models. This approach is based on the principle that combining weak models can result in a more robust and accurate model.

The algorithm works as follows: initially, a base model (e.g., a weak decision tree) is trained on the original data set. Then, weights are assigned to each data instance based on the accuracy of the base model in predicting those instances. In subsequent iterations, more weight is given to instances that were misclassified by the previous base model, allowing the next base model to focus on the more difficult to predict instances.

During the prediction stage, the results of all base models are combined by weighting, where the models that perform better on the training data set are given more weight in the final prediction.

AB is generally much more powerful and potent than other techniques by focusing on iteratively improving model performance through the combination of multiple weak models. Unlike traditional statistical methods, which may face difficulties in capturing nonlinear or complex relationships between variables, AB can adapt and continuously improve its predictive capability as more base models are added.

In terms of performance optimization, the most important hyperparameters in AdaBoostRegressor include the number of base models to use (n_estimators), the learning_rate that controls the contribution of each base model in the final combination, and the hyperparameters specific to the base models used in the boosting process. Adjusting these hyperparameters appropriately can help improve the generalization capability of the model and avoid overfitting.

2.4.3 BaggingRegressor (B)

Bagging Regressor [44–47] is another machine learning algorithm that is based on the concept of ensemble learning, specifically on the technique known as bagging (bootstrap aggregating). The fundamental idea is to build multiple independent predictive models and then combine their predictions to obtain a more robust and accurate final prediction, in a technique more or less similar to AB.

The algorithm works by creating multiple bootstrap samples (samples of equal size to the original data set, but with replacement) from the training data set. For each of these bootstrap samples, a separate regression model is trained, usually a base model such as a decision tree. Since each model is trained on a different bootstrap sample, they are inherently different from each other.

During the prediction stage, the regression predictions from each individual model are combined to produce a final prediction, usually by taking the average of all predictions.

Bagging Regressor stands out as a robust algorithm by taking advantage of the diversity of models trained on different bootstrap samples. This allows the final model to have lower variance and be less prone to overfitting compared to a single model trained on the entire dataset.

In terms of performance optimization, the most important hyperparameters in BaggingRegressor include the number of estimators (n_estimators), which determines how many models will be created in the ensemble, and the size of the bootstrap samples (max_samples), which controls the size of the samples used to train each individual model. Adjusting these hyperparameters appropriately can help improve the generalizability of the model and reduce the risk of overfitting.

2.4.4 SupportVectorRegressor (SVR)

SVR is a machine learning algorithm [48, 49] used for regression problems in particular. It is based on the idea of finding the optimal hyperplane that best separates data points in a high-dimensional space. In its methodology it focuses on finding a regression function that best fits the data while maintaining the widest possible margin of separation.

The algorithm transforms the input data into a high-dimensional feature space using a kernel function, such as linear, polynomial or radial kernel (RBF). In this high-dimensional feature space, SVR seeks to find the hyperplane that best fits the training data, where this hyperplane is defined by a set of support vectors, which are the data points closest to the hyperplane.

During training, SVR minimizes a loss function that penalizes both the discrepancy between model predictions and actual values and the violation of the separation margin. This is achieved using convex optimization techniques.

SVR outperforms traditional techniques based on robust statistics alone by being able to efficiently handle nonlinear and high dimensionality data through the kernel function. Unlike some traditional statistical methods that may face difficulties in modeling nonlinear relationships, SVR can effectively capture these relationships by transforming the data into a higher dimensional feature space.

In terms of performance optimization, the most important hyperparameters in SVR include the type of kernel to use (linear, polynomial, RBF), as well as kernel-specific hyperparameters such as the regularization coefficient (C) and the kernel coefficient (gamma). Properly adjusting these hyperparameters can help to improve the generalizability of the model and avoid overfitting.

2.4.5 BayesianRidge (BR)

BR [50, 51] is a regression algorithm that relies on the Bayesian framework to estimate model parameters. Actually, the algorithm uses inferential statistics, specifically the application of Bayes' theorem to infer the distribution of model parameters given the observed data.

The algorithm begins by modeling the relationship between the input variables and the target variable using a probability distribution. Instead of estimating a single set of values for the model parameters, Bayesian Ridge estimates a full distribution over these parameters, which allows capturing uncertainty in the estimates.

During training, Bayesian Ridge uses Bayesian methods to estimate the posterior distribution of the model parameters given the observed data and an initial prior distribution. This prior distribution can incorporate prior information about the model parameters, if available.

Once the posterior distribution is estimated, Bayesian Ridge can make predictions using the predictive distribution, which takes into account both the uncertainty in the model parameters and the inherent uncertainty in the input data.

Unlike some traditional statistical methods that can produce point estimates without accounting for uncertainty, Bayesian Ridge provides a full estimate of the posterior distribution, providing a more complete and realistic view of the regression problem.

In terms of performance optimization, the most important hyperparameters in Bayesian Ridge include the selection of the initial prior distribution for the model parameters, as well as regularization via an alpha precision parameter.

2.4.6 RidgeRegressor (R)

Ridge Regressor [52, 53] is a regression algorithm that is based on the ridge regression method, also known as ridge regression. The philosophy behind RidgeRegressor is to add a regularization penalty to the loss function of the standard linear regression model, with the goal of reducing overfitting and improving model generalization.

The algorithm works by minimizing a loss function consisting of two terms: the mean squared error (MSE) term that measures the discrepancy between model predictions and actual values, and a regularization penalty term that penalizes the magnitude of the model coefficients.

During training, RidgeRegressor adjusts the coefficients of the linear regression model such that it minimizes the weighted sum of these two terms. The regularization penalty controls the trade-off between fitting the training data accurately and keeping the model coefficients small to avoid overfitting.

RidgeRegressor outperforms traditional techniques based on robust statistics alone by incorporating the regularization penalty, which helps mitigate overfitting and improve the stability of model parameter estimates. RidgeRegressor can provide more stable and reliable estimates in such scenarios.

In terms of performance optimization, the most important hyperparameter in RidgeRegressor is the regularization parameter (alpha), which controls the strength of the penalty applied to the model coefficients.

2.4.7 ExtraTreesRegressor (ET)

ET [41] is a variant of the Random Forest algorithm that is characterized by its focus on the randomness and diversity of the constructed decision trees. Unlike Random Forest, which seeks to find the best threshold for each node split, ET selects the splitting thresholds randomly.

The algorithm works by constructing a "forest" of decision trees during training, similar to Random Forest. However, during the construction of each tree, ExtraTreesRegressor randomly selects a subset of features and split thresholds for each node, rather than exhaustively searching for the best thresholds.

This randomness in the selection of thresholds and features leads to greater diversity among the individual trees in the ensemble. As a result, Extra Trees Regressor tends to have greater variability in the individual trees, which can help reduce overfitting and improve the generalizability of the model.

In terms of performance optimization, the most important hyperparameters in Extra Trees Regressor include the number of trees in the "forest" (n_estimators), the maximum depth of each tree (max_depth), and the minimum number of samples required to split a node (min_samples_split).

2.4.8 GradientBoostingRegressor (GB)

GB is a machine learning algorithm [54, 55] that belongs to the boosting family of methods. Unlike other boosting algorithms that focus on improving the model by iteratively adding weaker models, GradientBoostingRegressor focuses on improving the model by iteratively adding stronger models that fit the residuals of the previous model.

The algorithm works sequentially, building a series of weak regression models, where each new model fits the residuals (differences between current model predictions and actual values). During training, GB fits the models in the direction of the downward gradient of the loss function, hence the name.

At each iteration, a new regression model is added to the ensemble, and adjusted by the gradient descent method to minimize the overall loss function. This technique allows GB to continuously improve model accuracy by adding additional models that focus on the residual errors of the previous model.

Unlike traditional methods that may have difficulty capturing nonlinear relationships or complex interactions between variables, GB can effectively accommodate high-dimensional data and nonlinearities.

In terms of performance optimization, the most important hyperparameters in GB include the number of base models to use (n_estimators), the learning rate (learning_rate) that controls the contribution of each base model in the final combination, and the maximum depth of each tree (max_depth).

2.4.9 HistGradientBoostingRegressor (HGB)

HGB [56, 57] is a variant of the Gradient Boosting algorithm that uses histograms to improve computational efficiency and model performance. Unlike GB, which uses decision trees based on complete features, HGB uses discrete histograms to represent the features, allowing for faster training and prediction speed.

The algorithm works by dividing the features into discrete intervals and constructing a histogram for each feature instead of evaluating all possible splits. This significantly reduces the amount of computation required during training and prediction, resulting in faster training times and more efficient models.

Compared to GB, this can be faster and more scalable on large datasets with many features. However, there may be a slight loss in model accuracy due to feature discretization.

In terms of performance optimization, HGB has similar hyperparameters to GB, such as the number of base models (n_estimators), the learning rate (learning_rate), and the maximum depth of each tree (max_depth). However, HGB also introduces additional hyperparameters related to feature discretization, such as the number of intervals (max_bins) and the number of bins per feature (max_bins_per_feature). Adjusting these hyperparameters appropriately can help improve the speed and accuracy of the model.

This algorithm offers an efficient and scalable alternative to GB, especially on large data sets, while providing comparable prediction results in terms of accuracy. The choice between the two algorithms depends on the specific needs of the problem, the amount of data and the importance of training and prediction speed.

2.4.10 TheilSenRegressor (TS)

TS [58] is a robust regression algorithm that is based on the Theil-Sen regression estimator. Unlike many other regression algorithms that rely on least-squares methods, S uses a robust estimation of the slope and the regression line intersection, which makes it less sensitive to outliers in the data.

The algorithm works by calculating all possible combinations of pairs of data points and estimating the slope and intersection of the regression line that best fits these combinations. Then, the median of all these slopes and the median of all the intersections are calculated to obtain the final slope and intersection estimates.

By focusing on the median of all possible slopes and intersections, TS is less sensitive to outliers or noisy data compared to traditional least squares methods, which can be significantly influenced by extreme values.

Unlike least squares methods, which can produce biased or unreliable estimates in the presence of outliers, TS provides a more stable and reliable estimate of the regression model parameters.

In terms of performance optimization, TS has no additional hyperparameters to adjust compared to other regression algorithms. Its robust approach and its ability to handle outliers without the need for additional hyperparameters make it attractive for applications where robustness is a priority. However, it can be slower on very large data sets due to its exhaustive approach to computing all possible combinations of pairs of data points.

2.5 Performance of models

Nine algorithms were applied to the data, and the performance of the imputation model was evaluated using two metrics specifically chosen for this case study.

The first metric used is the Kling-Gupta Efficiency skill score (KGEss), a variant of the Kling-Gupta Efficiency (KGE) metric developed by Gupta et al. [59]. KGEss assesses the degree of similarity or association between the target variable (e.g., observed series, x_{0i} for $i = 1, \dots, n$ data points) and the predicted variable (e.g., imputed series, x_{1i}) before and after the imputation process. It considers bias, variability, and correlation between the two time series, incorporating the coefficient of variation to address cross-correlation between bias and variability terms. KGEss is calculated using the formula:

$$\text{KGEss} = 1 - \frac{\sqrt{(r-1)^2 + (A-1)^2 + (B-1)^2}}{\sqrt{2}}, \quad -\infty \le \text{KGEss} \le 1, \tag{2}$$

where $A = CV_1/CV_0$, $B = \mu_1/\mu_0$, μ_1 and μ_0 are the means of the imputed and observed data, σ_1 and σ_0 are the standard deviations of the imputed and observed data, CV_1 and CV_2 are the coefficients of variation of the imputed and observed data, and r is the correlation coefficient between the imputed and observed data. A perfect match between the imputed and observed series results in a KGEss value of 1, indicating a high level of similarity. Lower values indicate increasing divergence between the two series, and a value close to or equal to zero suggests that the imputed series is no better than using the mean as an imputation method. A negative KGEss indicates that the observed series is a better estimator than the imputed series [60].

The second metric employed is the Percent Bias (PBIAS), which measures the average tendency of the imputed values to be higher or lower than the observed values. It is calculated as:

$$\text{PBIAS} = 100 \times \frac{\sum(x_{0i} - x_{1i})}{\sum(x_{0i})}, \tag{3}$$

where x_{0i} represents the original values in the dataset, and x_{1i} denotes the imputed values. The optimal value of PBIAS is 0, indicating accurate model imputation. Lower absolute values of PBIAS suggest better model performance. Positive values indicate an underestimation bias of the model, while negative values indicate an overestimation bias. In this study, |PBIAS| 15 is considered the convention for optimal model performance.

By utilizing these two metrics, the aim is to assess the quality of the imputation models and determine their effectiveness in recovering missing values in the physical, chemical, and biological composition of water samples for human consumption.

3 Results and discussions

From the correlation analysis between variables, it's calculated all the cross-correlation coefficients for each pair of variables in the dataset. The analysis reveals that the variable Total Alkalinity is highly positively correlated with Conductivity ($r = 0.630829$), Total Hardness ($r = 0.920800$), and Nitrites ($r = 0.626527$). Alkalinity, or the basicity of water, is a measure of its ability to neutralize acids. In natural waters, this property is primarily due to the presence of certain salts of weak acids, although the presence of weak and strong bases can also contribute. Generally, in natural waters, bicarbonates contribute the most to alkalinity, formed easily by the action of atmospheric carbon dioxide on the constituent materials of soils in the presence of water, as shown in the reaction: $CO_2 + CaCO_3 + H_2O \mapsto Ca^{2+} + 2HCO_3^-$.

A high presence of ions in water increases average conductivity and also affects water hardness due to the concentration of mineral compounds, particularly magnesium and calcium salts. While Alkalinity should have a greater impact on the measured pH, correlations in this dataset suggest that pH should also be correlated with Conductivity, Total Hardness, Alkalinity, and Nitrites. However, in terms of the data, this correlation has not emerged, or it is shown to be weak, with correlation coefficients of 0.255021, 0.219427, 0.262429, and 0.296015, respectively. Hardness also contributes strongly to sulfates ($r = 0.515953$).

Apparent color appears correlated with the variables Flavor ($r = 0.506351$) and Turbidity ($r = 0.851362$). Color in water bodies results from the presence of both suspended and dissolved substances and is referred to as apparent color. Therefore, it is entirely reasonable that its interference with these two variables is likely to be strong.

Total Organic Carbon is highly correlated with Iron content ($r = 0.76487$). Total Organic Carbon (TOC) refers to the carbon that is part of the organic substances in surface waters. In the contemporary environment, various natural and artificial substances contribute to increasing TOC levels. However, microorganisms can decompose this substance during the process of oxygen consumption. TOC generally originates naturally in plants and animals as a result of their metabolism, excretion, and decomposition. Additionally, effluents from industries using organic compounds serve as a significant source of TOC emissions into the environment. Companies discharging wastes into the sampled waters may also dispose of high iron content in the products derived from their industrial processes.

Among the strongest correlations found in this dataset is the relationship for Taste ~ (Apparent Color, $r = 0.506351$) + (Odor, $r = 0.997284$). Water taste and odor are organoleptic determinations with subjective assessments, lacking observation instruments, records, and units of measurement. These determinations are of particular interest in drinking water intended for human consumption. Waters acquire a salty taste from 300 ppm of Cl^- and a salty and bitter taste with more than 450 ppm of SO_4^{2-}. Free CO_2 gives it a pungent taste, while traces of phenols or other organic compounds can result in an unpleasant color and taste.

Figure 3 shows the dependence between some variables with maximum correlations found from the analysis of the original data set. Additionally, Fig. 4 shows the variability from some parameters by stations and year sample.

Looking at Fig. 4, it is evident that the variables and parameters analyzed in this study exhibit maximum or above-average values between the years 2011 and 2012 across most of the sampling stations and water collection points. This pattern suggests a potential association with the "La Niña" phenomenon observed during 2010–2011 [61–63]. The winter wave related to the 2010–2011 "La Niña" event is considered one of the most devastating natural disasters in Colombia's history, resulting in significant socio-economic losses. The term "La Niña" refers to the temporal and spatial distribution of oceanic-atmospheric indices during the 2010–2011 phenomenon [64, 65]. Within the spectrum of climate variability, the El Niño-Southern Oscillation (ENSO) phenomena, including the cold phase known as "La Niña" and the warm phase known as "El Niño", play a crucial role in shaping climate patterns across various regions of the Earth's surface. During "La Niña", sea level pressure tends to be lower in the western Pacific and higher in the eastern Pacific, while the opposite pattern is observed during "El Niño". This pressure field variation is known as the Southern Oscillation, with a standard measure being the difference in sea level pressure between Tahiti (18° S, 150° W) and Darwin (12° S, 131° E). Colombia is among the regions significantly affected by these phenomena. The presence of "La Niña" has had a noticeable impact on Colombia's climate, resulting in emergencies associated with slow flooding, flash floods, and landslides, leading to human and material losses.

The data imputation process was implemented using Python programming, and all calculations for the 10 evaluated models were executed on a laptop computer equipped with an Intel i5 core processor, 8GB of RAM, and the Ubuntu operating system, running within a Python Notebook. The entire task of constructing and imputing with the 10 models took approximately 3 min.

Throughout the imputation process, all variables are employed as inputs for imputing a specific variable. This approach implies that for the imputation of a parameter, no variables were excluded from being used as inputs. Instead, it is considered all variables as predictors for the imputation.

Table 3 provides a comprehensive summary of the evaluation metrics for all variables within the complete set of physico–chemical parameters of the water samples. The table presents the results for the assessed metrics, namely, the Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency skill score (KGEss), and Percent Bias (PBIAS).

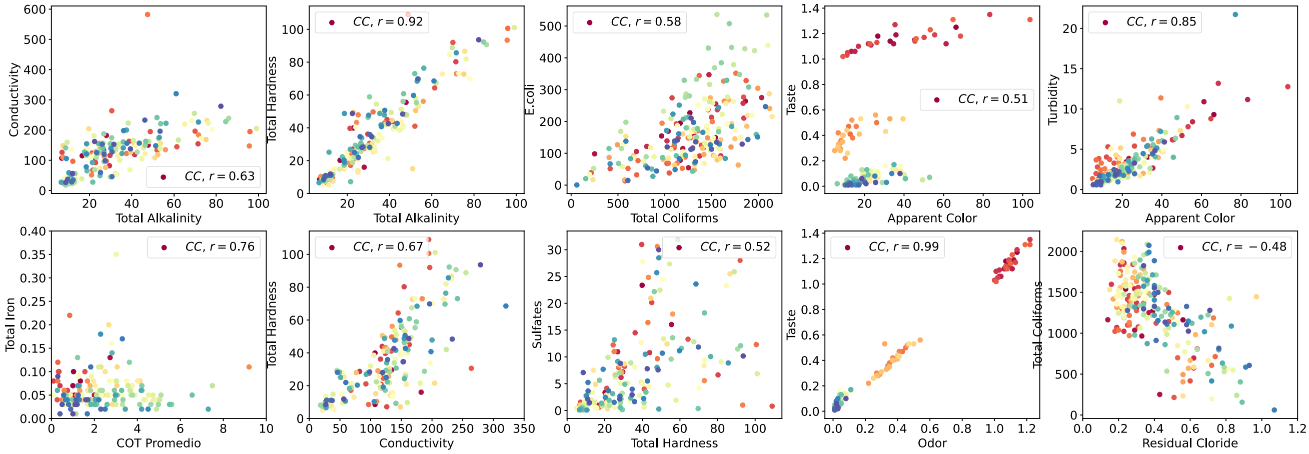

Fig. 3 Visual correlation between some variables from dataset studied. The scatter plot shown the mayor correlations ($|r| > 0.5$) for pair of parameters

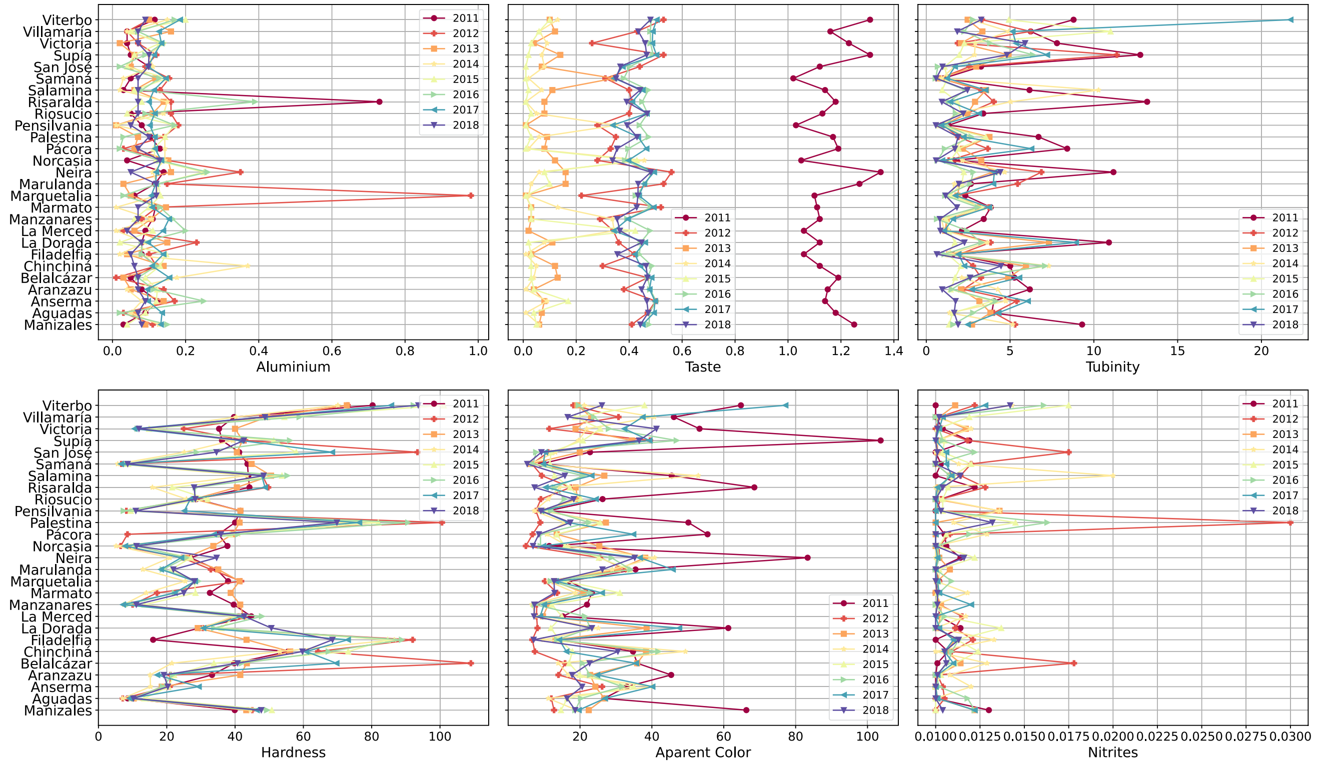

Fig. 4 Annual evolution of the values of some physico–chemical parameters throughout the seasons and years of sampling

Based on the previously mentioned imputation strategy, an imputed dataset have been generated using each implemented algorithm. In Fig. 5, present a visual representation of the overall impact of the strategy on the complete dataset for some ratio missing data at each variable, showcasing the average behavior of selected imputed variables characterized by varying degrees of missing data. It's essential to recognize that each algorithm comes with its set of advantages and drawbacks, potentially resulting in imputed values that either closely align with or deviate from the average values in the original imputed variable.

It is crucial to note that the proximity of imputed values to the average value in the original imputed variable doesn't inherently imply superiority or inferiority. This is primarily because information solely on the average imputed value per year lacks the context needed for evaluation. In reality, the distribution of values for each variable may undergo changes, and the loss of data can significantly impact and skew the distribution itself.

Based on the results presented in Table 3 and Figs. 5 and 6, it can be concluded that there is no single method or model that provides optimal estimation for all parameters of the water quality samples in the Colombian dataset. However, with the exception of certain parameters (such as maximum and average aluminum, maximum and average nitrates, maximum and average odor, and maximum and average taste) which do not yield good estimates based on the evaluation criteria, all other parameters can be imputed optimally by considering two evaluation metrics, namely KGEss and PBIAS.

This indicates that regardless of the metric used for evaluating the model, the measured variables can be effectively reproduced using the same model for both evaluation metrics. Among the tested models, the best performers according to the two metrics are Bayessian Ridge, Gradient Boosting, Ridge, Support Vector Machine, and Theil-Sen regressors, with the latter three exhibiting better overall performance.

These findings demonstrate the suitability of the selected models for imputing missing data in the water quality dataset, as they consistently yield favorable results based on the evaluation metrics.

By utilizing the best-performing imputation models, a new dataset can be constructed that contains the imputed values for the corresponding parameters. This allows for a fresh evaluation of the relationships between these parameters. In particular, four important relationships have been identified:

1. The relationship between Alkalinity Average and Hardness Average, Nitrites Average, Sulfates Average, and Conductivity Average exhibits a high coefficient of determination ($R^2 = 0.840$). The relationship can be expressed

Table 3 Best imputation models and corresponding goodness-of-fit indicator values per variable

Parameter/Model	AB	Bag	BR	ET	GB	HGB	R	RF	SVM	TS
Metric = KGEss										
Maximum Alkalinity	0.59	0.60	0.76	0.63	0.68	0.70	**0.78**	0.64	0.71	0.71
Alkalinity Average	0.50	0.49	0.56	0.51	0.49	0.48	0.56	0.52	**0.61**	0.57
Maximum Aluminium	0.11	−0.06	0.37	0.08	0.26	0.31	**0.48**	0.13	0.26	0.36
Aluminium Average	0.13	0.03	0.32	0.07	0.22	0.30	**0.46**	0.12	0.15	0.45
Cloruros Máximo	0.70	0.73	0.78	0.73	0.76	0.34	**0.83**	0.70	0.81	0.70
Clorides Average	0.71	0.75	0.81	0.74	0.77	0.38	**0.83**	0.72	0.82	0.79
Maximum TOC	0.94	0.94	0.85	0.94	0.94	0.05	0.82	0.94	**0.98**	0.92
TOC Average	0.87	0.86	0.87	0.87	0.88	0.47	0.88	0.87	**0.94**	0.90
Maximum Hardness	0.52	0.52	**0.72**	0.58	0.59	0.51	0.70	0.56	0.63	0.60
Hardness Average	0.55	0.54	0.63	0.57	0.55	0.46	0.62	0.55	**0.65**	0.58
Maximum Fluorides	0.64	0.48	0.71	0.54	**0.75**	0.34	0.72	0.72	0.73	0.71
Fluorides Average	0.56	0.51	0.63	0.53	**0.68**	0.19	0.69	0.55	0.45	0.59
Maximum Iron	0.73	0.61	**0.98**	0.76	0.83	0.18	0.17	0.68	0.85	0.69
Iron Average	0.61	0.62	0.76	0.62	0.65	0.36	0.85	0.60	0.43	**0.90**
Maximum Nitrites	0.00	0.09	0.31	0.15	0.12	0.18	0.32	0.10	−0.95	**0.44**
Nitrites Average	−6.19	−16.21	−2.47	−12.58	−12.09	−21.56	−2.07	−15.01	−16.80	**− 1.58**
Maximum Odor	−1.65	−1.88	0.15	−1.86	−1.69	−1.76	−0.25	−1.89	−1.57	**− 0.07**
Odor Average	−0.12	−0.17	0.28	−0.07	−0.05	−0.06	0.31	−0.15	**0.35**	−0.04
Maximum Taste	−1.48	−1.59	−1.37	−1.59	−1.47	−1.39	−0.96	−1.59	−1.11	**− 0.41**
Taste Average	−0.12	−0.16	0.11	−0.02	−0.03	−0.03	0.09	−0.16	**0.42**	−0.02
Maximum Sulfates	0.53	0.54	0.82	0.56	0.58	0.55	**0.78**	0.51	0.70	0.64
Sulfates Average	0.55	0.56	0.83	0.59	0.61	0.50	**0.84**	0.52	0.77	0.70
Metric = PBIAS										
Maximum Alkalinity	17.74	17.04	10.34	15.73	13.52	13.54	**9.63**	15.25	12.23	13.00
Alkalinity Average	17.11	17.26	14.92	16.55	17.50	19.42	14.79	16.52	**12.93**	14.49
Maximum Aluminium	37.68	46.30	29.45	38.35	32.49	36.72	25.23	36.72	31.39	30.31
Aluminium Average	37.01	43.06	31.42	39.60	33.90	38.05	27.85	37.60	36.18	33.42
Cloruros Máximo	16.39	14.68	12.10	15.15	13.31	38.67	**9.68**	16.25	9.93	16.19
Clorides Average	16.51	14.82	11.34	15.31	13.51	38.75	**10.63**	15.97	10.38	12.09
Maximum TOC	4.26	4.24	10.40	4.05	3.98	52.53	12.19	4.45	**1.68**	5.37
TOC Average	8.69	8.95	9.06	8.44	8.11	35.54	7.96	8.86	**3.96**	6.66
Maximum Hardness	18.05	18.43	**10.73**	15.58	14.97	19.08	11.29	16.31	13.27	14.86
Hardness Average	16.37	16.74	13.42	15.54	16.38	20.84	13.81	16.37	**12.49**	14.99
Maximum Fluorides	21.15	32.22	19.54	26.61	**15.18**	39.05	19.19	17.34	16.31	19.57
Fluorides Average	25.26	28.03	24.70	27.25	**19.70**	44.26	21.52	27.36	30.40	29.07
Maximum Iron	16.78	26.90	**− 0.71**	15.13	11.20	46.93	−61.88	20.76	9.37	19.65
Iron Average	21.23	21.72	14.41	20.95	19.45	37.33	10.24	22.62	30.24	**7.15**
Maximum Nitrites	42.45	41.31	34.13	38.89	42.02	40.00	35.50	40.86	73.95	37.71
Nitrites Average	88.84	88.39	89.09	88.49	88.52	88.89	89.23	88.60	93.43	89.17
Maximum Odor	53.15	52.25	40.46	50.78	52.15	51.69	42.35	52.91	42.11	49.81
Odor Average	45.27	46.24	36.83	43.45	43.46	44.16	36.49	45.74	30.74	**0.94**
Maximum Taste	48.86	48.74	39.56	47.58	48.79	48.25	40.03	49.19	35.96	42.67
Taste Average	45.06	45.71	38.48	41.59	42.35	42.75	39.17	45.34	27.95	**11.62**
Maximum Sulfates	22.76	22.84	9.59	21.22	20.57	22.97	**12.44**	23.84	14.45	18.72
Sulfates Average	21.95	22.35	9.43	19.99	19.66	25.82	**10.19**	23.73	11.53	15.65

Bold values indicate to the best performance escenario for variable and machine learning method used, respectively

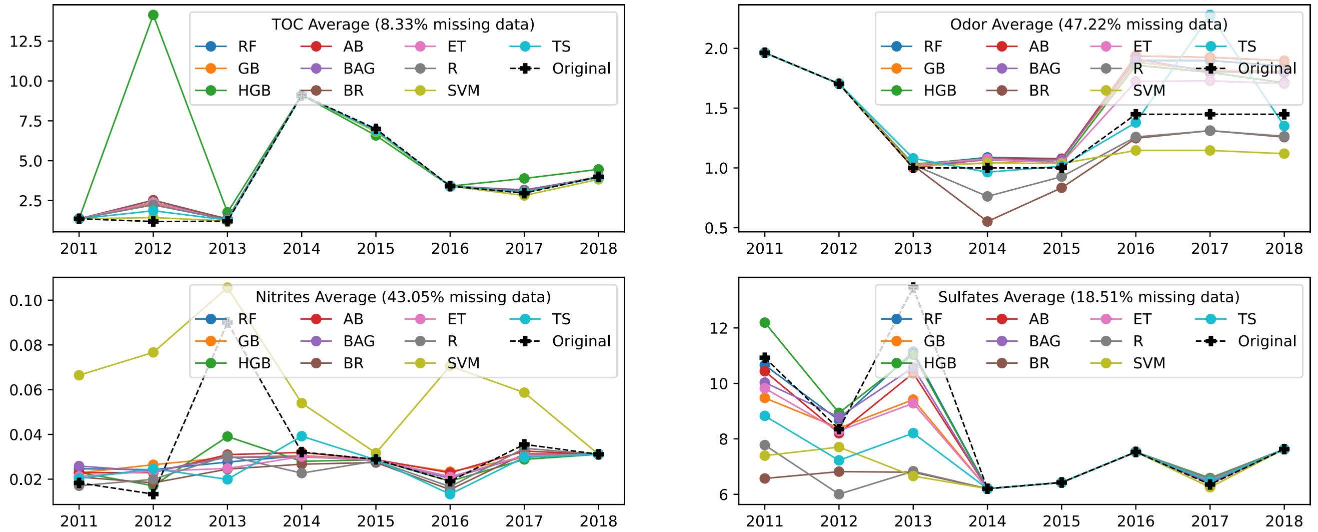

Fig. 5 The average value of some imputed variables (TOC, Odor, Nitrites and Sulfates Averages) is shown in reference to the year in which they were taken for the various machine learning algorithms implemented

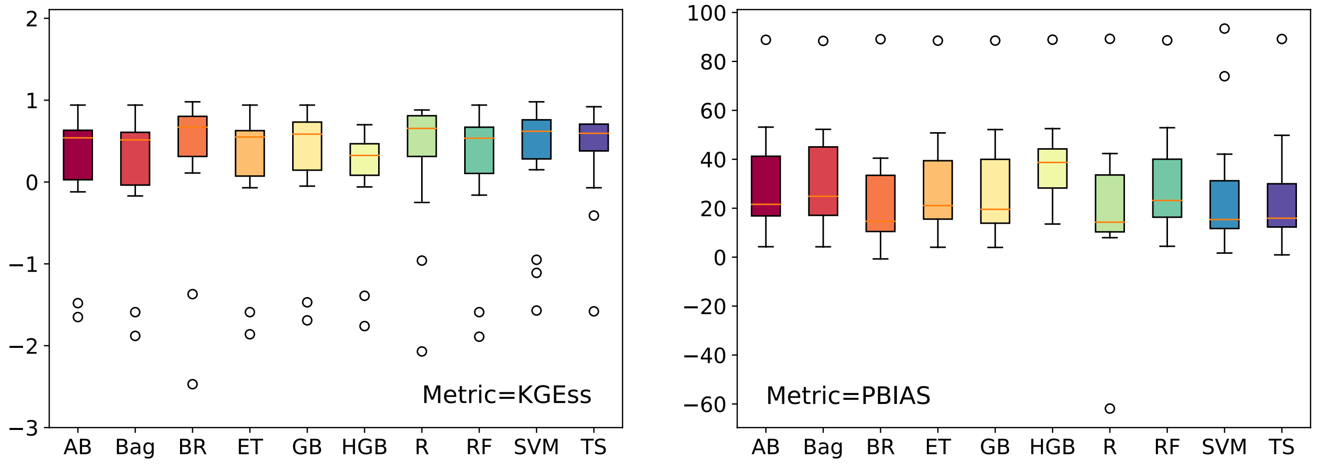

Fig. 6 Distribution of the measured values for the evaluation metrics in each of the models due to the imputed parameters

as: Alkalinity Average = 3.2125 + (0.7751) × (Hardness Average) + (−120.4619) × (Nitrites Average) + (−0.3692) ×(Sulfates Average) + (0.0355) × (Conductivity Average).

2. The TOC content is found to be related to the levels of Nitrites Average and Total Iron. The relationship is represented by the equation: TOC Average = 2.7807 + (81.7279) × (Iron Average) + (−329.3762) × (Nitrites Average), with a coefficient of determination of $R^2 = 0.540$.
3. The subjective parameters, such as Odor Average, Apparent Color Average, and Turbidity Average, demonstrate a significant relationship with certain chemical and organic properties. The relationship can be expressed as: Odor Average = 0.0140 + (0.9536) × (Taste Average) + (−0.0008) × (Aparent Color Average) + (−0.0010) ×(Turbinity Average), with a high coefficient of determination of $R^2 = 0.985$.

These relationships highlight the dependencies and interactions between different water quality parameters, providing valuable insights into the dataset and improving our understanding of the underlying processes.

Finally, a model for predicting the water quality index in terms of various parameters exhibits a high correlation with a high coefficient of determination of $R^2 = 0.834$:

$$
\begin{aligned}
\mathrm{IRCA}^{(imputed)} =&29.5554 + (-0.0076) \times (\text{Alkalinity Average})\\
&(-0.0581) \times (\text{Hardness Average})\\
&+ (0.0395) \times (\text{Sulfates Average})\\
&(-13.0703) \times (\text{Total Residual Cloride})\\
&+ (-0.0166) \times (\text{E.Coli Average})\\
&(0.0271) \times (\text{Coliforms Average}).
\end{aligned} \tag{4}
$$

This scenario for a statistical regression model for the prediction of the water quality index using the imputed variables can be compared with another scenario in which the same regression has been done using the original data. For this case:

$$
\begin{aligned}
\mathrm{IRCA}^{(no\_imputed)} =&97.2615 + (0.0236) \times (\text{Alkalinity Average})\\
&(-0.0285) \times (\text{Hardness Average})\\
&+ (0.0602) \times (\text{Sulfates Average})\\
&(-0.4076) \times (\text{Total Residual Cloride})\\
&+ (-0.0018) \times (\text{E.Coli Average})\\
&(0.0012) \times (\text{Coliforms Average}),
\end{aligned} \tag{5}
$$

achieving a coefficient of determination of $R^2 = 0.095$. In this case the number of instances that have been used for the development of the last regression has been 80% of the number of instances in the regression with imputed values due to the missing values contained in the data.

For the calculation of the coefficient of determination, it is understood, of course, that this statistic will be calculated for the degree of relationship between observed and predicted variables in the relationships previously found. In the case of the first three relationships, the data corresponding to the data that have been imputed are used, so that there is a one-to-one correspondence between the predicted and predicted related data. For the case of the IRCA calculation in terms of the global predictor variables the same applies as above, however in the case of the relationship with the original non-imputed data the relationship is obtained by removing from the dataset (not taking into account) the variables with missing data.

4 Conclusions

This study analyzed the performance of ten imputation methods for biological, chemical, and physical parameters in surface water samples intended for human consumption. The objective was to observe their impact on water quality indices for surface samples in reservoirs and streams. While it is always preferable to minimize non-response in surveys, the reality is that many parameters are often not measured due to various reasons.

The study found that imputation techniques based on field-collected information generated the best results and preserved data distribution. Machine Learning approaches, with their ability to minimize error and randomness, proved to be effective. The use of bootstrapping techniques and multiple estimators further improved predictions.

The analysis was performed without an a priori choice of imputation method, focusing on generating robust estimators that satisfy water quality indicator predictions. None of the methods considered the sample design structure. The objective was to minimize distortion in the distribution of intervening variables.

The results demonstrated that by considering 56 water quality measurement parameters and employing ten machine learning-based imputation methodologies, it was possible to predict missing values accurately. The imputed data also improved the determination of empirical relationships. For example, the relationship between Alkalinity and variables such as Hardness, Nitrite content, Electrical Conductivity, and average Sulfate content exhibited a high coefficient of determination ($R^2 = 0.840$). The local water quality index prediction in Colombia also showed a high coefficient of determination ($R^2 = 0.834$) using the most important parameters.

Based on the evaluation metrics, Bayesian Ridge, Gradient Boosting, Ridge, Support Vector Machine, and Theil-Sen regressors were identified as the most efficient algorithms, producing better results in terms of PBIAS and KGEss. While no method was ideal for all parameters, these algorithms demonstrated precise estimation for individual parameters related to water quality.

This approach differs from a previous study [23] in several ways. Firstly, boosting variables or helper parameters were not used for imputation but instead, imputation was performed over the entire dataset. Secondly, in addition to the five algorithms tested in the previous study, five more algorithms using ensemble methods and linear regression were evaluated, demonstrating the superiority of the newly considered algorithms. Finally, this study utilized additional information on the maximum measurement and number of samples for each parameter, which helped to improve the imputation process and reduce data variance.

Overall, the findings highlight the effectiveness of the proposed approach in imputing missing data and improving the estimation of water quality parameters.

Acknowledgements The author is grateful to J.Lobatón for the valuable comments and for the careful work of reviewing the conceptualization, the writing of the manuscript and revising language issues. The author would also like to thank the Universidad Tecnológica de Bolivar for their unconditional support throughout the research process. Finally, I thank the Secretaría de Ambiente del Departamento de Caldas, Colombia, for sharing the data used in this research. The author thank the manuscript reviewers for their invaluable contributions through insightful comments and suggestions. Their input has played a crucial role in enhancing the quality and depth of our article, propelling it towards greater academic excellence. Their dedication and expertise have been essential for the continuous improvement of this work, and we are profoundly grateful for their time and commitment to our scientific advancement.

Author contributions D.S-P designed the model and the computational framework and analyzed the data, carried out the implementation, performed the calculations, and wrote and reviewed the whole manuscript.

Funding No funds have been received for the development of this research.

Data availability For purposes of replication and sharing methodologies, a dataset that has been used for the generation of the manuscript: can be found at: Sierra Porta, David (2022), "Dataset: Efficient improvement for water quality analysis with large amount of missing data", Mendeley Data, V1, doi: 10.17632/8y42cbc7h8.1.

Declarations

Competing interests The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. El-Dessouky HT, Ettouney HM. Fundamentals of salt water desalination, vol. 1. Amsterdam: Elsevier; 2002. p. 669. https://doi.org/10.1016/B978-0-444-50810-2.X5000-3.
2. Al-Karaghouli A, Kazmerski LL. Energy consumption and water production cost of conventional and renewable-energy-powered desalination processes. Renew Sustain Energy Rev. 2013;24:343–56. https://doi.org/10.1016/j.rser.2012.12.064.
3. Sharqawy MH, Lienhard JH, Zubair SM. Thermophysical properties of seawater: a review of existing correlations and data. Desalination Water Treat. 2010;16(1–3):354–80. https://doi.org/10.5004/dwt.2010.1079.
4. Páll E, Niculae M, Kiss T, Şandru CD, Spînu M. Human impact on the microbiological water quality of the rivers. J Med Microbiol. 2013;62(Pt 11):1635. https://doi.org/10.1099/jmm.0.055749-0.
5. Issaka S, Ashraf MA. Impact of soil erosion and degradation on water quality: a review. Geol Ecol Landsc. 2017;1(1):1–11. https://doi.org/10.1080/24749508.2017.1301053.
6. Heathwaite A. Multiple stressors on water availability at global to catchment scales: understanding human impact on nutrient cycles to protect water quality and water availability in the long term. Freshw Biol. 2010;55:241–57. https://doi.org/10.1111/j.1365-2427.2009.02368.x.
7. Ferreira CS, Walsh RP, Ferreira AJ. Degradation in urban areas. Curr Opin Environ Sci Health. 2018;5:19–25. https://doi.org/10.1016/j.coesh.2018.04.001.
8. Novotny V. Water quality: diffuse pollution and watershed management. Hoboken, New Jersey: John Wiley & Sons; 2002.
9. Chaudhry FN, Malik M. Factors affecting water pollution: a review. J Ecosyst Ecogr. 2017;7(225):1–3. https://doi.org/10.4172/2157-7625.1000225.
10. Zhang Y-F, Fitch P, Thorburn PJ. Predicting the trend of dissolved oxygen based on the kpca-rnn model. Water. 2020;12(2):585. https://doi.org/10.3390/w12020585.

11. Zhang Y, Thorburn PJ. Handling missing data in near real-time environmental monitoring: a system and a review of selected methods. Future Gener Comput Syst. 2022;128:63–72. https://doi.org/10.1016/j.future.2021.09.033.
12. Osman MS, Abu-Mahfouz AM, Page PR. A survey on data imputation techniques: water distribution system as a use case. IEEE Access. 2018;6:63279–91. https://doi.org/10.1109/ACCESS.2018.2877269.
13. Chiu PC, Selamat A, Krejcar O, Kuok KK, Herrera-Viedma E, Fenza G. Imputation of rainfall data using the sine cosine function fitting neural network. Int J Interact Multimedia Artif Intell. 2021. https://doi.org/10.9781/ijimai.2021.08.013.
14. Zhang Y-F, Thorburn PJ, Xiang W, Fitch P. Ssim–a deep learning approach for recovering missing time series sensor data. IEEE Internet Things J. 2019;6(4):6618–28. https://doi.org/10.1109/JIOT.2019.2909038.
15. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402. https://doi.org/10.4097/kjae.2013.64.5.402.
16. Soley-Bori M. Dealing with missing data: Key assumptions and methods for applied analysis, 2013. https://www.math.wsu.edu/faculty/xchen/stat115/lectureNotes3/Marina%20Dealing%20with%20missing%20data.pdf
17. Tabari H, Hosseinzadeh Talaee P. Reconstruction of river water quality missing data using artificial neural networks. Water Qual Res J Canada. 2015;50(4):326–35. https://doi.org/10.2166/wqrjc.2015.044.
18. Tang J, Deng C, Huang G-B. Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst. 2015;27(4):809–21. https://doi.org/10.1109/TNNLS.2015.2424995.
19. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmos Environ. 1998;32(14–15):2627–36. https://doi.org/10.1016/S1352-2310(97)00447-0.
20. Gutmann H-M. A radial basis function method for global optimization. J Glob Optim. 2001;19(3):201–27. https://doi.org/10.1023/A:1011255519438.
21. Ghosh J, Nag A. An overview of radial basis function networks. Radial basis function networks 2: new advances in design, 2001;1–36, https://doi.org/10.1007/978-3-7908-1826-0_1.
22. Srebotnjak T, Carr G, Sherbinin A, Rickwood C. A global water quality index and hot-deck imputation of missing data. Ecol Indicat. 2012;17:108–19. https://doi.org/10.1016/j.ecolind.2011.04.023.
23. Rodríguez R, Pastorini M, Etcheverry L, Chreties C, Fossati M, Castro A, Gorgoglione A. Water-quality data imputation with a high percentage of missing values: a machine learning approach. Sustainability. 2021;13(11):6318. https://doi.org/10.3390/su13116318.
24. Sierra-Porta D. Hydrogeochemical evaluation of water quality suitable for human consumption and comparative interpretation for water quality index studies. Environ Process. 2020;7(2):579–96. https://doi.org/10.1007/s40710-020-00426-7.
25. Ball RO, Church RL. Water quality indexing and scoring. J Environ Eng Div. 1980;106(4):757–71. https://doi.org/10.1061/JEEGAV.0001067.
26. Lumb A, Sharma T, Bibeault J-F. A review of genesis and evolution of water quality index (wqi) and some future directions. Water Qual Expo Health. 2011;3(1):11–24. https://doi.org/10.1007/s12403-011-0040-0.
27. Noori R, Berndtsson R, Hosseinzadeh M, Adamowski JF, Abyaneh MR. A critical review on the application of the national sanitation foundation water quality index. Environ Pollut. 2019;244:575–87. https://doi.org/10.1016/j.envpol.2018.10.076.
28. Brown RM, McClelland NI, Deininger RA, Tozer RG. A water quality index-do we dare. Water and sewage works, 1970;117(10).
29. Dinius S. Design of an index of water quality 1. JAWRA J Am Water Resourc Assoc. 1987;23(5):833–43. https://doi.org/10.1111/j.1752-1688.1987.tb02959.x.
30. Barros JC. Aplicação do Índice de Qualidade das Águas (IQA-CETESB) no açude Gavião para determinação futura do Índice de Qualidade das Águas Brutas para fins de Abastecimento Público (IAP), 2012. https://propi.ifto.edu.br/ocs/index.php/connepi/vii/paper/viewFile/2850/2313
31. Boyacioglu H. Development of a water quality index based on a European classification scheme. Water Sa, 2007. https://doi.org/10.4314/wsa.v33i1.47882.
32. Banda TD, Kumarasamy M. Development of a universal water quality index (uwqi) for South African river catchments. Water. 2020;12(6):1534. https://doi.org/10.3390/w12061534.
33. Hurley T, Sadiq R, Mazumder A. Adaptation and evaluation of the Canadian council of ministers of the environment water quality index (ccme wqi) for use as an effective tool to characterize drinking source water quality. Water Res. 2012;46(11):3544–52. https://doi.org/10.1016/j.watres.2012.03.061.
34. Khan AA, Paterson R, Khan H. Modification and application of the Canadian council of ministers of the environment water quality index (ccme wqi) for the communication of drinking water quality data in newfoundland and labrador. Water Qual Res J. 2004;39(3):285–93. https://doi.org/10.2166/wqrj.2004.039.
35. Cash K, Wright R. Canadian Water Quality Guidelines for the Protection of Aquatic Life. CCME, 2001. https://prrd.bc.ca/wp-content/uploads/post/prrd-water-quality-database-and-analysis/WQI-Technical-Report-en.pdf
36. Ocampo-Duque W, Ferre-Huguet N, Domingo JL, Schuhmacher M. Assessing water quality in rivers with fuzzy inference systems: a case study. Environ Int. 2006;32(6):733–42. https://doi.org/10.1016/j.envint.2006.03.009.
37. Van Rossum G, Drake FL Jr. Python reference manual. Amsterdam: Centrum voor Wiskunde en Informatica Amsterdam; 1995.
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
39. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013;108–122.
40. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
41. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1.
42. Drucker H. Improving regressors using boosting techniques, 1997. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d8226a52ebc70c8d97ccae10a74e1b0a3908ec1.
43. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39. https://doi.org/10.1006/jcss.1997.1504.
44. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40. https://doi.org/10.1007/BF00058655.

45. Breiman L. Pasting small votes for classification in large databases and on-line. Mach Learn. 1999;36(1):85–103. https://doi.org/10.1023/A:1007563306331.
46. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44. https://doi.org/10.1109/34.709601.
47. Louppe G, Geurts P. Ensembles on random patches. 2012. https://doi.org/10.1007/978-3-642-33460-3_28.
48. Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):1–27. https://doi.org/10.1145/1961189.1961199.
49. Platt J, et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif. 1999;10(3):61–74.
50. Tipping ME. Sparse bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1(Jun):211–44.
51. MacKay DJ. Bayesian interpolation. Neural Comput. 1992;4(3):415–47. https://doi.org/10.1162/neco.1992.4.3.415.
52. McDonald GC. Ridge regression. Wiley Interdiscip Rev Comput Stat. 2009;1(1):93–100. https://doi.org/10.1002/wics.14.
53. Wieringen WN. Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169, 2015.
54. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. vol. 2. New York: Springer. 2009.
55. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;1189–1232.
56. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30.
57. Zhang H, Si S, Hsieh C-J. Gpu-acceleration for large-scale tree boosting, 2017. arXiv preprint https://doi.org/10.48550/arXiv.1706.08359, arXiv:1706.08359.
58. Dang X, Peng H, Wang X, Zhang H. Theil-sen estimators in a multiple linear regression model. Olemiss Edu, 2008.
59. Gupta HV, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. J Hydrol. 2009;377(1–2):80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
60. Koskinen M, Tahvanainen T, Sarkkola S, Menberu MW, Laurén A, Sallantaus T, Marttila H, Ronkanen A-K, Parviainen M, Tolvanen A, et al. Restoration of nutrient-rich forestry-drained peatlands poses a risk for high exports of dissolved organic carbon, nitrogen, and phosphorus. Sci Total Environ. 2017;586:858–69. https://doi.org/10.1016/j.scitotenv.2017.02.065.
61. Seiler LM, Fernandes EHL, Martins F, Abreu PC. Evaluation of hydrologic influence on water quality variation in a coastal lagoon through numerical modeling. Ecol Model. 2015;314:44–61. https://doi.org/10.1016/j.ecolmodel.2015.07.021.
62. Murshed MF, Aslam Z, Lewis R, Chow C, Wang D, Drikas M, Leeuwen J. Changes in the quality of river water before, during and after a major flood event associated with a la niña cycle and treatment for drinking purposes. J Environ Sci. 2014;26(10):1985–93. https://doi.org/10.1016/j.jes.2014.08.001.
63. Boening C, Willis JK, Landerer FW, Nerem RS, Fasullo J. The 2011 la niña: So strong, the oceans fell. Geophys Res Lett. 2012. https://doi.org/10.1029/2012GL053055.
64. Hoyos N, Escobar J, Restrepo J, Arango A, Ortiz J. Impact of the 2010–2011 la niña phenomenon in Colombia, South America: the human toll of an extreme weather event. Appl Geogr. 2013;39:16–25. https://doi.org/10.1016/j.apgeog.2012.11.018.
65. Restrepo JD, Kettner AJ, Syvitski JP. Recent deforestation causes rapid increase in river sediment load in the Colombian Andes. Anthropocene. 2015;10:13–28. https://doi.org/10.1016/j.ancene.2015.09.001.