Archives of Environmental Protection
Vol. 52 no. 1 pp. 3–15

PL ISSN 2083-4772
DOI: 10.24425/aep.2026.158379

Preparation of water quality dataset using outlier detection and imputation methods

Mehmet Kazim YETİK

Karabük University, Turkey

Corresponding author`s e-mail: kazimyetik@karabuk.edu.tr

Keywords: water quality, outlier detection, missing data imputation, boxplot, normalization, Mean Squared Error (MSE).

Abstract: Reliable water quality assessment depends on high-quality datasets; however, datasets obtained from field measurements often contain outliers and missing values, which directly affect subsequent analytical processes. This study presents a practical system application that enhances dataset integrity and reliability for advanced artificial intelligence analyses. The proposed approach integrates exploratory visualization (boxplots) for outlier detection, statistical outlier treatment, mean- and regression-based estimation for missing values, and normalization for ensuring comparability. The application utilizes a river water quality dataset from Türkiye, including Cl^-, Fe, K^+, Na^+, SO_4^{2-}, TKN, TN, and turbidity parameters. Model performance demonstrated strong results, with Mean Squared Error (MSE) values for the identified outliers ranging from 0.00002 (turbidity) to 0.135 (K^+). Comparative analysis of raw and post-processed datasets revealed that systematic outlier handling and targeted imputation improved data consistency and reduced modeling uncertainty, thereby enabling more reliable ANN- and GA-based predictive modeling. The proposed methodological framework is practical, reproducible, and easily integrable into water quality monitoring systems, supporting data-driven management and policy decision-making.

Introduction

In recent years, growing concerns over environmental pollution and the sustainable management of water resources have highlighted the need for accurate monitoring and analysis of water quality parameters. Water quality directly influences ecosystem health and public well-being, making its assessment a critical component of environmental management strategies. However, reliable analysis requires high-quality datasets, which are often compromised by measurement errors, missing values, and outliers–common issues in field-based data collection.

In modeling studies and time series analyses, the accuracy and reliability of predictive outcomes are strongly dependent on the quality of input data. Consequently, data preprocessing, particularly outlier detection and missing data imputation, has become an essential step in the preparation of robust analytical datasets. Without these processes, erroneous or incomplete data can significantly distort model outcomes and hinder effective decision-making.

Among the various techniques employed in data preprocessing, **outlier detection** and **missing data imputation** are particularly crucial. A well-prepared dataset is a prerequisite for obtaining precise and meaningful analytical results. This is especially relevant in studies of **water quality parameters**, where both data collection challenges and advanced analytical processes can significantly influence the reliability of findings.

In several countries, water quality assessment requires manual sampling and subsequent laboratory analysis (Islam Khan et al., 2022). To facilitate decision-making, policymakers and stakeholders frequently rely on the **Water Quality Index (WQI)**, a widely used metric for assessing and communicating the pollution status of water bodies (Tripathi & Singal, 2019; Sadiq et al., 2022). Numerous researchers have developed numerical methods to aid in the planning and/or design of water quality studies, including Roger A. Falconer (1992) and Addico G. et al., (2006). Given the significance of water quality assessment, improving the integrity and completeness of datasets through effective data preprocessing methods is essential for generating reliable conclusions in environmental studies.

Studying water quality parameters in rivers can be challenging due to field-related limitations. In addition, measurement errors or deviations may occur on-site due to meteorological factors, such as air temperature, rainfall, and other environmental conditions. Since multiple factors can affect the measured values, various approaches are used to identify and eliminate erroneous data from the data set.

Outliers are observations that are significantly different from other data in the data set and can negatively affect the analysis results. Therefore, it is necessary to accurately detect and manage outliers. The boxplot is an effective visual tool for this purpose, as it displays the distribution of data and highlights outliers. By determining the median, quartiles and outliers, boxplots enable the rapid identification of, facilitating the creation of accurate and clean datasets in model studies. Missing data is another common issue in datasets, which can jeopardize the accuracy of analyses. Missing data imputation involves estimating and filling these missing values. Regression-based imputation is one of the effective methods used for this purpose. In agricultural datasets, for example, regression imputation allows for more accurate evaluation of relationships among climate, soil properties, and plant development parameters. By reducing the impact of missing data, this approach enhances the performance and reliability of advanced analyses.

Water quality is of critical importance because it has a direct impact on environmental sustainability and human health. Pollution of water resources and effects of climate change are affected by various industrial activities and agricultural practices. In this context, accurate analysis of water quality parameters is necessary for the development of water management and protection strategies. However, abnormal values (outliers) and missing data, which occur during data collection, can eliminate the reliability of analyses and the possibility of obtaining accurate results.

This study focuses on a critical issue in environmental data analysis: identifying erroneous measurements and approximating missing values within large water quality datasets. Specifically, it demonstrates the applicability of statistical data preprocessing techniques to a river water quality dataset collected in Türkiye, emphasizing that the elimination of biases significantly enhances the reliability of analytical outcomes. The proposed methodological framework presented in this study has been designed to be universally applicable to river systems worldwide, illustrating its global relevance and adaptability. Key parameters analyzed in the dataset include Cl^-, Fe, K^+, Na^+, SO_4^{2-}, TKN, TN, and turbidity. The study evaluates the effectiveness of preprocessing methods, including outlier detection via box plots and missing value estimation through regression analysis, in improving data accuracy and consistency. Comparative assessment between raw and processed datasets reveals that preprocessing enhances data integrity and provides a solid foundation for robust predictive modeling, including approaches based on artificial intelligence (AI), artificial neural networks (ANN), and genetic algorithms (GA).

The integration of these preprocessing steps contributes to the creation of reliable and valid datasets, thereby facilitating the development of advanced analytical and decision-support tools for water quality and basin management. Furthermore, the study systematically examines the theoretical and practical feasibility of preprocessing techniques for correcting incomplete or erroneous data, ultimately supporting the advancement of sustainable water management strategies across river systems (Isaac & Siddiqui, 2022). Consequently, the proposed framework may serve as a reference model for global river monitoring and management initiatives.

The main objectives of this research are to (i) demonstrate the practical implementation of statistical preprocessing techniques for improving the quality of riverine datasets, (ii) evaluate the effects of these techniques on analytical accuracy and model performance, and (iii) establish a reproducible methodological framework that can be adapted to diverse hydrological and environmental datasets worldwide.

Materials and Methods

Study Area (Data Collection)

Station No: 13-23-00-048; Station Name: ARAS CAYI-KARIT
This study investigates the removal of inaccurate data and the handling of missing values using data cleaning and imputation techniques. By applying relevant data processing techniques, the study aims to produce a complete dataset free of erroneous values. Accordingly, analyses were performed using five-year water quality data from the Araç River.

Within the scope of the research, water quality parameters in the Araç River were monitored and analyzed by the General Directorate of State Hydraulic Works (DSI). Data were collected on a monthly basis between 2011 and 2014. Field measurements were carried out by DSI experts at a monitoring station operated by the General Directorate of State Hydraulic Works. The geographical coordinates of the station are 33.024216° longitude and 41.231477° latitude, and the station location is presented in Figure 1.

The geographical location of the Araç River and the monitoring station are illustrated in Figure 2. The Araç River constitutes a significant hydrological system, as it flows through multiple provinces and is referred to by different names in various regions. Ultimately, it discharges into the Black Sea under the name 'Yenice River,' serving as a vital freshwater source that supports diverse ecosystems along its course. Given its ecological importance, continuous monitoring and evaluation of water quality parameters are essential. In this context, numerous observation stations operated by the General Directorate of State Hydraulic Works (DSI) are situated along the river, enabling regular data collection for effective water resource management.

Methodology

Water quality parameters in Turkey are monitored through on-site analyses and laboratory testing. River samples are collected at regular intervals by DSI experts; some analyses are carried out in the field, while others are subjected to detailed tests in accredited laboratories. More than 50 water quality parameters are monitored, with analyses carried out every three months, resulting in the creation of a long-term data set. In this study, basic water quality parameters such as chloride ion (Cl^-), Iron (Fe), potassium ion (K^+), sodium ion (Na^+), sulphate ion (SO_4^{2-}), Total Kjeldahl nitrogen (TKN), total nitrogen (TN) and turbidity are evaluated using five-year data.

Statistical summaries such as mean (μ), maximum, minimum, difference, variance (s^2) and coefficient of variation (CV) for these parameters are presented in Table 1 (eqs 1-4).

$$Difference = \frac{(Max - Min)}{Min} \cdot 100 \qquad (1)$$

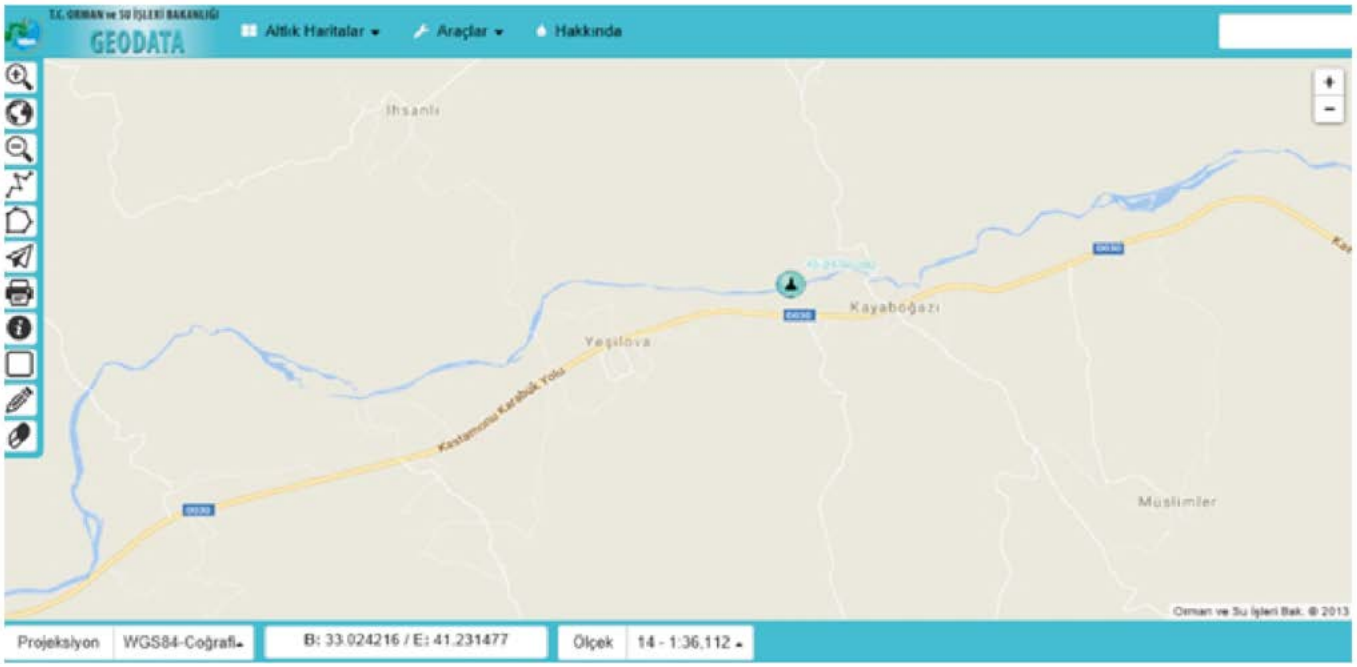

Fig. 1. Location of the station on the Araç River

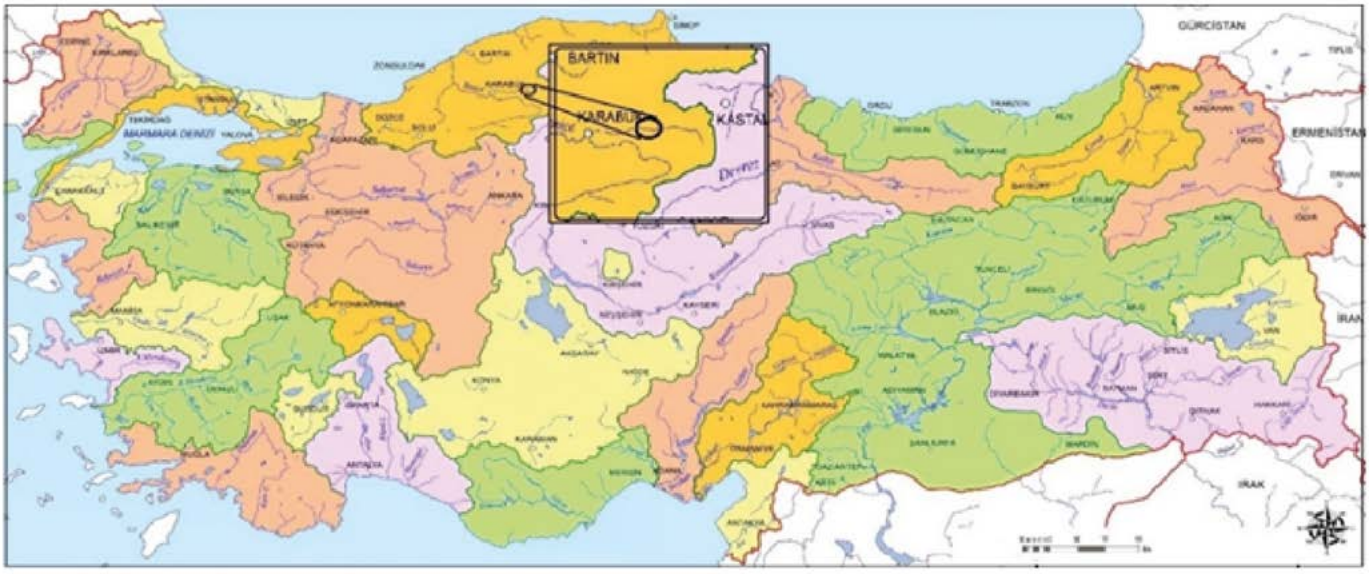

Fig. 2. Location of the sampling station within the basin, shown on the map of Turkey

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2)$$

Standard Deviation (SD)

$$a = \sqrt{\frac{1}{n}\sum_{i=1}^{n} '(x_1 - \propto)^2} \qquad (3)$$

$$CV = \frac{a}{\propto} = \qquad (4)$$

The primary aim of this study was to organize and refine the dataset based on selected water quality parameters obtained from analyses conducted by experts of the State Hydraulic Works (DSİ). Initially, statistical analyses were performed on the raw values of these parameters, and the corresponding results are presented in Table 1. During this process, outliers and missing data were identified and are summarized in Table 2, which details their distribution across parameters and time periods. To ensure data integrity and analytical reliability, various preprocessing methods were applied to the dataset. Owing to data confidentiality, the dataset was visualized in processed form rather than in its entirety. All analyses in this study were carried out using actual measurement data to maintain the accuracy and representativeness of the observed water quality conditions.

Table 1 Statistical Summary of Raw Water Quality Data.

Unit	Parameters	Mean	max	min	Difference %	Var	CV
mg/L	Cl-	5,60	9,67	2,03	376,35	5,01	0,40
µg/L	Fe	607,22	1408,00	151,10	831,83	143283,42	0,62
mg/L	K+	1,66	1,97	1,35	45,93	0,03	0,11
mg/L	Na+	13,48	22,07	6,51	239,01	20,60	0,34
mg/L	SO4=	37,87	65,00	17,50	271,43	200,94	0,37
mg/L	TKN	0,28	0,64	0,10	540,00	0,03	0,62
mg/L	TN	0,65	1,05	0,46	128,26	0,03	0,28

Notes:
- Cl-: Chloride Ion (Cl),
- Fe: Iron,
- K+: Potassium ion,
- Na+: Sodium Ion,
- SO4: Sulphate Ion
- TKN: Total Kjeldahl Nitrogen,
- TN: Total Nitrogen
- Turb: Measurement of water clarity.

Table 2: Table of missing and outlier values in selected parameters

Parameters	2010-3	2010-9	2011-9	2011-11	2012-5	2012-9	2013-5	2013-9	2013-11	2014-3	2014-5	2014-9	2014-11
Cl-	+++	+++	+++	+++	||||||||	+++	+++	+++	+++	+++	+++	+++	+++
Fe	+++	+++	||||||||	||||||||	+++	+++	+++	+++	+++	+++	||||||||	+++	XXXX
K+	+++	XXXX	+++	+++	+++	||||||||	+++	+++	+++	+++	+++	XXXX	+++
Na+	+++	+++	+++	+++	+++	+++	||||||||	+++	+++	+++	+++	+++	+++
SO4=	+++	+++	+++	+++	+++	+++	+++	+++	||||||||	+++	+++	+++	+++
TKN	||||||||	||||||||	+++	+++	+++	+++	+++	+++	+++	+++	+++	||||||||	+++
TN	+++	||||||||	+++	+++	+++	+++	XXXX	+++	+++	+++	+++	+++	+++
Turb	+++	+++	+++	||||||||	+++	+++	+++	+++	+++	+++	+++	+++	+++

\|\|\|\|\|\|\|\| - missing data XXXX - Outliers

To address data irregularities and enhance dataset consistency, several preprocessing techniques were employed, including outlier detection using box plot analysis, normalization of parameter values to minimize scale disparities, and regression-based imputation to estimate missing measurements. Collectively, these methods improved the statistical robustness of the dataset and established a reliable foundation for subsequent modeling and interpretation of river water quality dynamics.

The detection of outliers was first comprehensively defined by Tukey and has since become a fundamental method in statistical data analysis. Despite advancements in statistical theory and computational technologies, the box plot method remains widely used due to its simplicity and effectiveness in identifying extreme values. Its applicability across diverse disciplines has been well documented. For instance, Ahmad et al. (2001) applied the box plot method during the preprocessing stage of surface water quality data analysis.

Today, many statistical software packages and data analysis tools incorporate the box plot method as an integrated component of outlier detection process. Contemporary software such as MATLAB, Microsoft Excel, and Minitab effectively utilize this method for both visual and statistical identification of outliers in datasets (Wu et al., 2020; Jancosek & Pajdla, 2014; Garlits et al., 2023, Guo et al., 2015). These applications demonstrate that Tukey's approach not only provides a theoretical foundation but also retains its relevance as a reliable, simple, and widely applicable tool in modern data analysis.

Missing data pose significant challenges to both the statistical analysis of water quality datasets and the performance of machine learning models. In this context, Betrie et al., (2016) and Zhang & Thorburn, (2022) evaluated the effectiveness of various imputation techniques for handling missing values. In general, machine learning studies employ a range of imputation methods to mitigate the adverse effects of missing data issues (Yang, 2022).

Preliminary analyses play a crucial role in uncovering relationships and interdependencies among parameters. These analyses enable the identification of key variables that influence one another, revealing patterns and interactions essential for accurate modeling and interpretation of water quality data. Through visualization of correlations and dependencies within the dataset, valuable insights into the underlying data structure are obtained. Figure 3 illustrates these interactions, thereby enhancing the understanding of variable relationships.

Outlier analysis

Outlier analysis is a statistical technique used to identify values that significantly deviate from the overall distribution

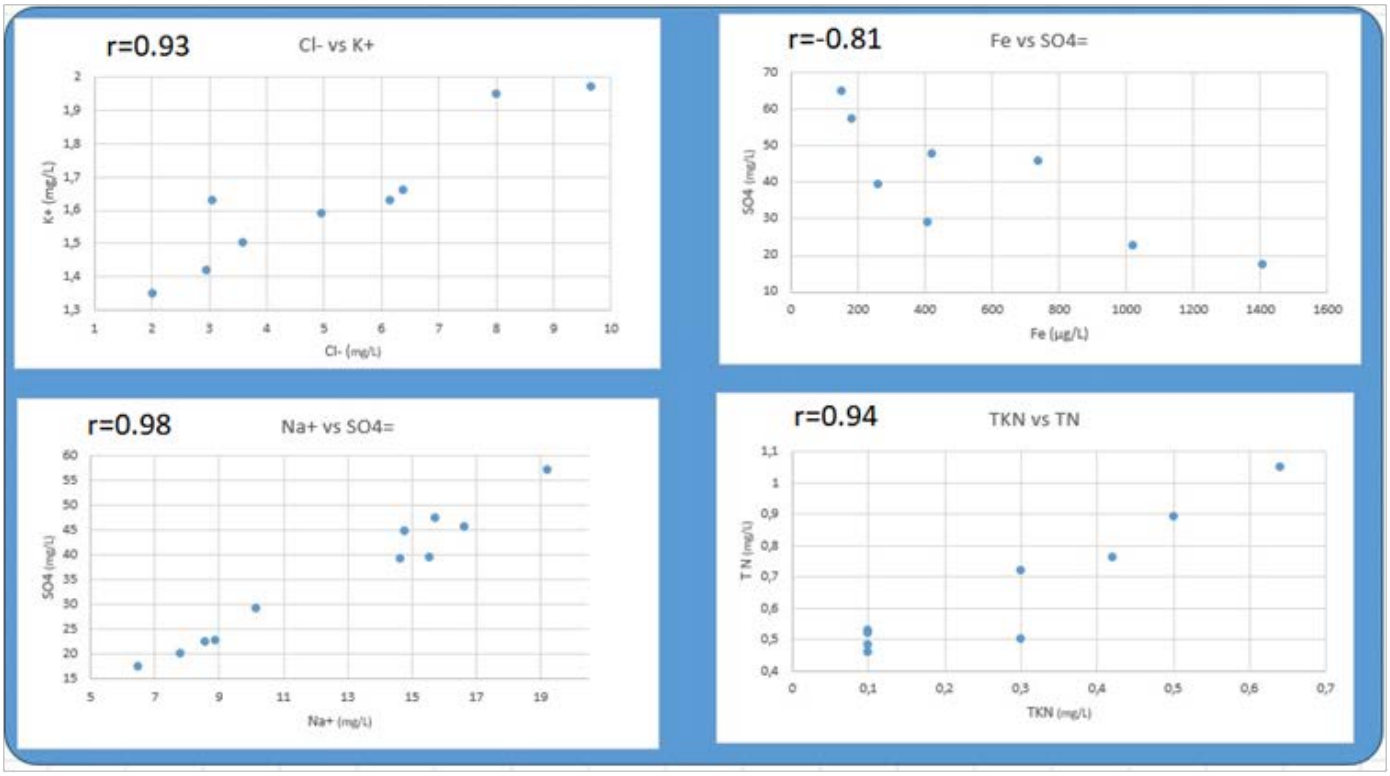

Fig. 3: The dot plot charts of the parameters with high correlations

of a dataset. It plays a particularly important role in fields such as statistical modeling and data mining, where analytical accuracy depends heavily on data quality. Outliers may arise from various sources, including data entry or measurement errors, systematic issues, or genuine phenomena representing meaningful deviations. Therefore, accurate identification and appropriate treatment of outliers are essential for enhancing the reliability of analytical results. Nevertheless, outliers should be carefully evaluated before removal, as they may, in some cases, contain valuable information that contributes to system understanding or the discovery of new patterns. The scientific methods used in outlier analysis are described below.

a. Statistical Methods: The following techniques are based on the fundamental statistical characteristics of the data:

Z-Score (Standard Score): This method quantifies the distance of a data point from the mean in units of standard deviation, or distance. The Z-score is calculated using the following equation (Eq. 5):

$$Z = \frac{X - \mu}{\sigma} \qquad (5)$$

where x represents the observation, μ is the dataset mean, and σ denotes the standard deviation. A high absolute Z-score indicates that a data point may be considered an outlier. In practice, values with |Z|>3 are commonly classified as outliers (Tripathi & Singal, 2019; Li et al., 2021).

Grubbs test is a hypothesis-based method designed to detect a single outlier in a normally distributed dataset. It is particularly suitable for small sample sizes (Wilrich, 2013).

The Dixon Q test is another statistical method used to identify outliers, especially in small datasets (Efstathiou, 2006).

b. Machine Learning Methods

Machine learning methods aim to identify anomalies in data through learning-based techniques. In contrast to statistical approaches, these methods are capable of detecting more complex and flexible patterns by analyzing the structural characteristics of the data (Maniruzzaman et al., 2018).

- K-Means Clustering: The K-Means algorithm is a clustering technique that identifies outliers based on group distances. This method partitions the dataset into a predefined number of clusters based on the distances between them. Each data point is assigned to the closest cluster center, and the algorithm iteratively updates the cluster centers to optimize the grouping. Data points located at large distances from their respective cluster centers, or those that do not fit well within any cluster, are considered potential outliers. Therefore, by calculating the distances between data points and cluster centers, the K-Means technique is a useful tool for identifying outliers.
- Support Vector Machines (SVM) are powerful machine learning methods widely used for outlier detection, especially through the "One-Class SVM" model. Using data from a single class, One-Class SVM learns the boundary that encloses normal observations by mapping data points into a high-dimensional feature space and constructing a separating hyperplane. Observations that lie outside this boundary or are located far from the hyperplane are classified as outliers. This approach is especially effective for large and complex datasets and has demonstrated strong performance in anomaly detection tasks.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:** DBSCAN is a density-based

clustering algorithm that identifies clusters by grouping data points in dense regions. The algorithm relies on two parameters: the radius (epsilon) and the minimum number of points that are supplied. Data points that do not meet the density criteria and do not belong to any cluster are considered outliers. Unlike conventional clustering methods, this approach can identify clusters of varying densities and forms, which makes it particularly useful for large and unevenly distributed datasets.

- Isolation Forest and Outlier Detection: Isolation Forest is a tree-based machine learning algorithm specially designed for outlier detection. The method isolates observations by randomly partitioning the dataset using randomly selected features and split values. Since outliers are sparse in the data distribution, they tend to be isolated with fewer splits, resulting in shorter paths. Because of its low processing cost, this method is commonly chosen for large-scale anomaly detection problems and offers high efficiency in large datasets.

c. **Distribution and Distance-Based Methods**

Distribution and distance-based methods rely on the distances between data points and the underlying properties of data distribution for outlier detection. Liu (2001) emphasized that these approaches are particularly effective in identifying anomalies based on density differences. By analyzing local data densities, these methods classify observations located in low-density regions or at significantly greater distances from other data points as outliers. Specifically, local density-based techniques enable anomaly detection by computing a density score for each data point relative to its neighbors. Due to their sensitivity to data structure, these methods provide reliable results, especially for high-dimensional and complex datasets.

Mahalanobis Distance is suitable for finding outliers in multivariate data sets.

Based on the distance calculated according to the covariance structure of the data points.

Euclidean Distance measures the straight-line distance between data points; data points far from the cluster are considered potential outliers.

Local Outlier Factor (LOF) method evaluates the degree to which data point derives from its local neighborhood by comparing its density to that of its nearest neighbor.

d. Visualization-Based Methods

Visualization-based methods detect anomalies by examining the locations of outliers within the data distribution. By representing data in two or three dimensions, these techniques facilitate the identification of outliers. Common approaches include scatter plots, box plots, and principal component analysis (PCA), which allow for the visual recognition of unusual points in the dataset. These methods provide preliminary insights that complement statistical and machine learning-based approaches, thereby enhancing the overall effectiveness of anomaly detection procedures (Xie et al., 2020).

- Scatter Plot:
  - Scatter plots examine the relationship between two variables and allow for the visual identification of extreme points.
- Histograms and Density Plots:
  - Histograms and density plots are fundamental tools in exploratory data analysis for visualizing the distribution of a dataset. A histogram shows the frequency of data points within predefined intervals, while a density plot provides a smoothed estimate of the probability distribution. Potential outliers appear as extreme values that deviate significantly from the main distribution, manifesting as isolated bars in histograms or low-density regions in density plots,. These visualizations are valuable for assessing data quality and informing subsequent preprocessing steps.
- Boxplot:
  Boxplots provide a simple and effective way to visually detect outliers using the interquartile range (IQR) method. Key elements of a boxplot include:
  - **Left/Lower Boundary of the Box**: Q1 (1st Quartile)
  - **Right/Upper Boundary of the Box**: Q3 (3rd Quartile)
  - **Line Inside the Box**: Median (2nd Quartile)
  - **Whiskers (Line Extensions)**: Values within the range of Q1 - 1.5 × IQR and Q3 + 1.5 × IQR.
  - **Outliers**: Data points smaller than Q1 - 1.5 × IQR or larger than Q3 + 1.5 × IQR are displayed as small dots outside the whiskers.

Boxplot Calculation Method

Several studies investigating water quality parameters have employed variance and the coefficient of variation (CV) as key statistical indicators to assess data variability (Azhar et al., 2015). As presented in Table 1, certain parameters, such as BOD_5 and turbidity (Turb), exhibit maximum and minimum values that deviate considerably from the mean, resulting in relatively high variance values. The boxplot method has also been widely applied in outlier detection to identify such deviations. Numerous studies have utilized boxplot visualizations to reduce data noise and improve analytical accuracy (Islam Khan et al., 2022, Dawson, 2011). This approach is particularly effective in detecting extreme values in water quality parameters, enabling the correction and refinement of measurements through the removal or adjustment of outliers (Tripathi & Singal, 2019, Horvat et al., 2021). The methodologies adopted in this study are discussed in detail in the following sections.

Outlier detection methods are generally categorized into statistical approaches, distance-based techniques, machine learning-based models, and visualization methods. Among these, boxplots serve as an effective visualization tool for detecting and analyzing outliers. In boxplots, the interquartile range (IQR) is represented by a box encompassing the central 50% of the data, defined by the first (Q1) and third (Q3) quartiles. The median is indicated within the box, while whiskers extend from Q1 to the minimum value and from Q3 to the maximum value. Outliers are usually identified by data points that deviate significantly from the distribution of the dataset as a whole and are located outside of the whiskers.

In this study, boxplots were used to detect outliers,which helped reduce data noise and maintain dataset integrity. Fig. 4 illustrates the process of constructing boxplots and identifying outliers. Whiskers typically extend beyond the box itself, but in some cases, they may be shorter, representing a more uniform distribution with well-defined boundaries (Dawson, 2011). To improve comparability and aggression of results, the boxplots were constructed using normalized data, calculated as follows (eq. 6).

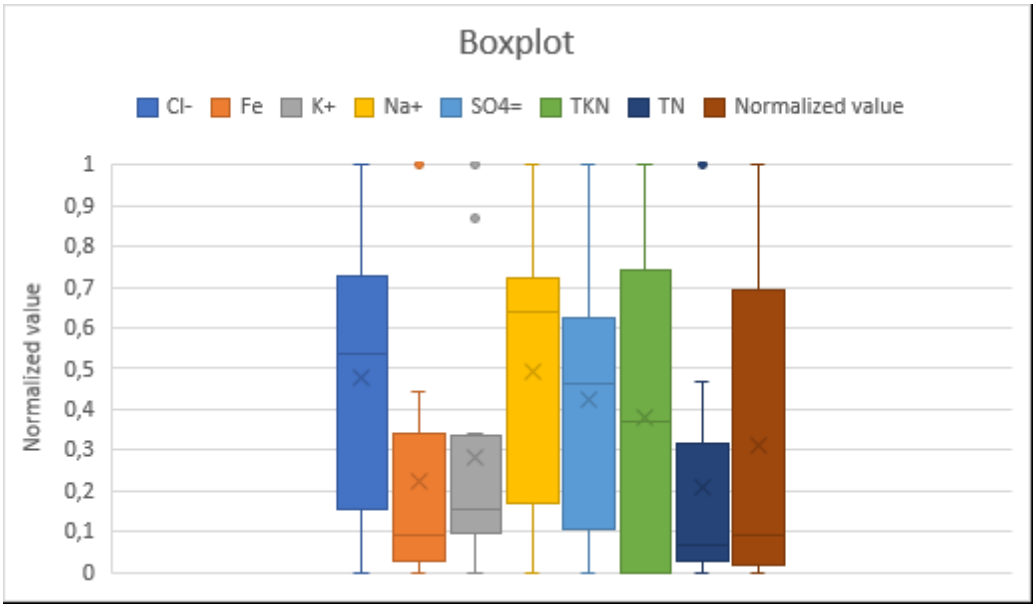

Fig. 4. For outlier detection, boxplot graphics (with normalized values) were used for chloride ion (Cl), iron (Fe), potassium ion (K+), sodium ion (Na+), sulphate ion (SO_4^{2-}), TKN, total nitrogen (TN), and turbidity (Turb) For outlier detection, boxplot graphics (with normalized values) were used for chloride ion (Cl), iron (Fe), potassium ion (K+), sodium ion (Na+), sulphate ion (SO_4^{2-}), TKN, total nitrogen (TN), and turbidity (Turb)

$$X_{Normalized} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \qquad (6)$$

Normalization also enhanced the reliability of the preprocessing stage by enabling a more consistent interpretation of outliers. The results of the analysis, including outlier identification, are presented in Figure 4.

As observed in the boxplot graphs, certain data points, represented by the colored symbol ¤ , fall outside the whiskers extending from the first quartile (Q1) to the minimum and from the third quartile (Q3) to the maximum. Additionally, data points marked with an asterisk (*) are located at a considerable distance from the central distribution within the box, indicating potential outliers. Specifically, iron (Fe), potassium (K^+), and total nitrogen (TN) exhibit notable deviations from the overall dataset structure, suggesting the presence of outliers. The remaining parameters lie within threshold limits and therefore do not require further processing. Table 3 summarizes the boxplot values for the selected parameters. In the figure above, the outlier values identified using the boxplot method are visualized, whereas the corresponding numerical values are provided in the table below. To enable collective visualization of all boxplots, normalized values are displayed in the figure; however, the actual extreme values are reported in the table for reference.

Table 3 Outlier values and boxplot analysis values obtained as a result of boxplot analysis for outlier value detection.

Parameter	Q1	Q3	Median	Bottom	Up
Fe	225.38	1116.25	414.9	151.1	1408
K+	1.52	1.63	1.97	1.35	1.97
TN.	0.49	0.55	0.86	0.46	1.05

Visualization-based methods aim to identify anomalies by using data visualization techniques, making outlier detection more intuitive and interpretable. The approaches proposed by Xie et al., (2020) emphasize the effectiveness of visual analysis tools for high-dimensional and complex datasets. By visually representing the characteristics and distributions of the data points, these methods facilitate the identification of regions where anomalies deviate from the main data structure. Commonly employed techniques include scatter plots, parallel coordinates, and dimensionality reduction methods such as t-distributed stochastic neighbor embedding (t-SNE). These visualization tools enable rapid identification of anomalies within the data, thereby supporting analytical processes and strengthening decision-making.

Missing Value Methods in Data Analysis

Handling missing values is a critical step in data preprocessing, as incomplete datasets can adversely affect the performance of statistical analyses, machine learning models, and other modeling algorithms. Missing data may arise from a variety of factors, including sensor malfunctions, human errors, or technical limitations during data collection. To address missing values effectively, a range of methods can be employed, which are generally categorized into deletion methods, imputation techniques, and advanced machine learning approaches.

Imputation

In the following section, several imputation methods are examined individually to address missing data within the dataset. Each method is discussed in terms of its underlying statistical principles, applicability to water quality parameters,

and potential impact on the accuracy and reliability of the analytical results.

a. ***Deletion Methods****:* Deletion methods represent the simplest approach to handling missing data and are generally effective when the proportion of missing values is small and the missingness occurs at random.
   - **Listwise Deletion**: This method removes any row containing at least one missing value. Although straightforward to implement, listwise deletion can result in substantial information loss and reduced statistical power when a large number of rows contain missing data (Dong & Peng, 2013; Salgado et al., 2016).
   - **Pairwise Deletion**: Instead of excluding entire rows, pairwise deletion uses all available data for each individual analysis. While this approach preserves more data than listwise deletion, it may complicate interpretation due to varying sample sizes across different statistical calculations (Dong & Peng, 2013, Salgado et al., 2016).

b. ***Simple Imputation Methods:*** Simple imputation methods are commonly used to handle missing data by replacing missing values with estimates derived from the observed data. These approaches aim to preserve dataset completeness while minimizing the impact of missing values on statistical analysis. Typical techniques involve substituting missing values with measures of central tendency, such as the mean, median, or mode, depending on the data type. Although simple imputation methods are computationally efficient and easy to implement, they may introduce bias or underestimate data variability when the missingness mechanism is not random. Consequently, these methods are most appropriate for datasets with a low proportion of missing values or when the missing data are assumed to be randomly distributed (Kadengye et al., 2012).
   - **Mean/Median Imputation**: Missing values are replaced with the mean or median of the observed values for the corresponding variable. Although simple and computationally efficient, this approach can distort variance and weaken relationships between variables.
   - **Mode Imputation**: Commonly applied to categorical variables, this method replaces missing values with the most frequently occurring category in the dataset.
   - **Forward/Backward Fill**: Commonly used in time-series datasets, this technique imputes missing values using the most recent previous observation (forward fill) or the subsequent observation (backward fill).

c. ***Advanced Imputation Methods:*** Advanced imputation methods aim to provide more accurate and reliable estimates for missing values by exploiting relationships and dependencies between variables in the dataset. For example, multiple imputation models the uncertainty of missing values by generating several plausible datasets, which are then combined to produce robust statistical inferences. Similarly, the K-Nearest Neighbors (KNN) imputation method identifies the most similar data points based on proximity in the feature space to predict missing values. These methods are particularly effective for datasets with complex patterns or high-dimensional data. However, they may require additional computational resources and careful parameter tuning to avoid overfitting or noise (Fouad et al., 2021).
   - **K-Nearest Neighbors (KNN) Imputation**: Replaces a missing value with the average (or weighted average) of values from the nearest neighbors based on feature similarity.
   - **Multiple Imputation**: Creates multiple plausible datasets by imputing missing values several times and combines results to account for uncertainty in imputed values.
   - Regression Imputation: Predicting Missing Values

Missing data is a common issue encountered in any data analysis process. These gaps may arise from various causes, such as data collection errors, unanswered survey questions, device malfunctions, or data loss. Missing values can influence analysis results and complicate accurate modeling. Therefore, appropriately predicting missing data is a crucial step in data science and statistical analysis. Regression imputation is a powerful and frequently used method for predicting missing values (Holt & Benfer, 2000).

The Basic Principle of Regression: Regression analysis aims to model the relationship between a dependent variable (target variable) and one or more independent variables. By defining this relationship, estimates can be generated for unknown or missing values. Regression analysis is commonly used for imputing missing data, particularly in datasets containing numerical variables.

Using Regression to Predict Missing Data: The application of regression methods to predict missing data is usually carried out in the following steps:

1. Data Review and Relationship Analysis: The first step is to identify which variables contain missing data. Next, the relationships between the missing variable and other variables in the dataset are examined. If the missing values can be explained by one or more observed variables, these variables can be used as independent predictors. In essence, the model predicts missing values by leveraging information from related variables.
2. Model Selection and Training: Linear regression is one of the most commonly used models for predicting missing values and may be appropriate when a linear relationship between variables is assumed. In this step, the **regression** model is specified based on the identified independent variables.
3. Applying the Model to the Training Set: The regression model is trained using observations that do not contain missing values. Model parameters (e.g., regression coefficients) are estimated using this complete subset of the training data.
4. Prediction and Filling of Missing Values: The trained model is then used to estimate missing values. These predicted values may be compared against known data (where available) to assess model performance. This validation process enhances the reliability of the imputation, resulting in more accurate and consistent datasets.

Advantages of Filling Missing Values with Regression (D. Yang et al., 2023)

- High Accuracy,
- Reducing Data,
- Data Set Completion.

Difficulties of Filling Missing Values with Regression,

- **Effect of Outliers**: Outliers in a regression model can adversely affect parameter estimation. Therefore, outliers may need to be identified and addressed during model training.
- **Selection of Independent Variables**: Selecting appropriate independent variables is essential for accurately estimating missing values. The inclusion of irrelevant or inappropriate variables can lead to biased or inaccurate estimates.
- **Assumption of Linearity**: In linear regression models, a linear relationship between the dependent and independent variables is assumed. If this assumption is violated, the model's estimation accuracy may be reduced.

The imputation methods described above were applied individually to the dataset, and correlations between each imputed version and the original dataset were examined. Among the tested approaches, the Regression Imputation was identified as the most suitable technique, as it resulted in the smallest deviation from the observed data values. Consequently, missing values in the dataset were imputed using the Regression Imputation to preserve the statistical integrity and representativeness of the data.

Results

After outliers were identified using box plot analysis (summary statistics are presented in Table 3), the corresponding parameters were further examined by comparison with data from other monitoring station along the same river. These extreme values were found to be inconsistent with seasonal variations of the river, therefore they were removed from the dataset. Subsequently, missing value analyses were performed to replace the removed values in the data set (Betrie et al., 2016; Zhang & Thorburn, 2022). However, since all the data for a particular month were missing, no imputation was performed for that period, and the data for that month were entirely excluded from the analysis.

Various approaches for handling missing data involve replacing missing values with the mean or median of the relevant dataset. Some studies suggest that the use arbitrary values may be acceptable in specific contexts (Yang, 2022). Among these approaches, linear regression based on the least squares method has been shown to be particularly effective for imputing missing data (Wang et al., 2012). In this method, correlations among parameters are evaluated, and a linear prediction model is constructed using the variable with the highest correlation to estimate missing values. Table 4 shows the correlation coefficients among the parameters in the dataset. Estimated values were then computed in accordance with the positions of the missing data.

A strong positive correlation (r = 0.93) was observed between Cl- and K+, indicating that these parameters tend to increase concurrently. Similarly, a strong correlation was identified between Fe and SO4 (r = -0.81), while SO4= and Na+ exhibited a very strong positive correlation (r = 0.98), suggesting that these ions may be affected by similar sources or geochemical processes. In contrast, a strong negative correlation (r = -0.79) between Cl- and turbidity (Turb) indicates an inverse relationship between these variables. However, regression-based imputation of Turbidity produced inconsistent trends in the dataset, indicating that the model did not adequately capture the underlying relationship. A value compatible with the overall data pattern was determined using a trial-and-error approach, with total Kjeldahl nitrogen (TKN) selected as the reference parameter due to its more stable relationship with Turbidity.

In general, the significant positive and negative correlations presented in the table provide valuable insights into the sources of the parameters and their sensitivity to environmental processes. These relationships are important factors to consider when developing water quality management and prediction models.

Regression analysis is a powerful and widely used method for estimating missing data. Appropriate model selection and careful identification of independent variables help ensure accurate estimation of missing values. However, since regression analysis relies on certain assumptions and has inherent limitations, thorough data preprocessing and model validation are required to improve the accuracy of the model. Overall, this approach provides an effective means of addressing missing data, thereby improving the reliability and robustness of data analysis processes.

Missing parameter estimation was performed using linear regression between the two parameters exhibiting the highest correlation in the table (Chen et al., 2022). Using this approach, deficiencies in the dataset were eliminated (Burchard-Levine et al., 2014). The imputation model was constructed based on the

Table 4: Correlation coefficients among parameters in the raw dataset (after outlier detection).

Parameters	Cl-	Fe	K+	Na+	SO4=	TKN	TN	Turb
Cl-		-0,61	0,93	0,86	0,56	-0,73	-0,40	-0,79
Fe	-0,61		-0,62	-0,78	-0,81	-0,71	-0,43	0,32
K+	0,93	-0,62		0,77	0,36	-0,54	-0,18	-0,57
Na+	0,86	-0,78	0,77		0,98	-0,21	-0,13	-0,76
SO4=	0,56	-0,81	0,36	0,98		0,04	-0,27	-0,31
TKN	-0,73	-0,71	-0,54	-0,21	0,04		0,94	0,69
TN	-0,40	-0,43	-0,18	-0,13	-0,27	0,94		0,48
Turb	-0,79	0,32	-0,57	-0,76	-0,31	0,69	0,48	

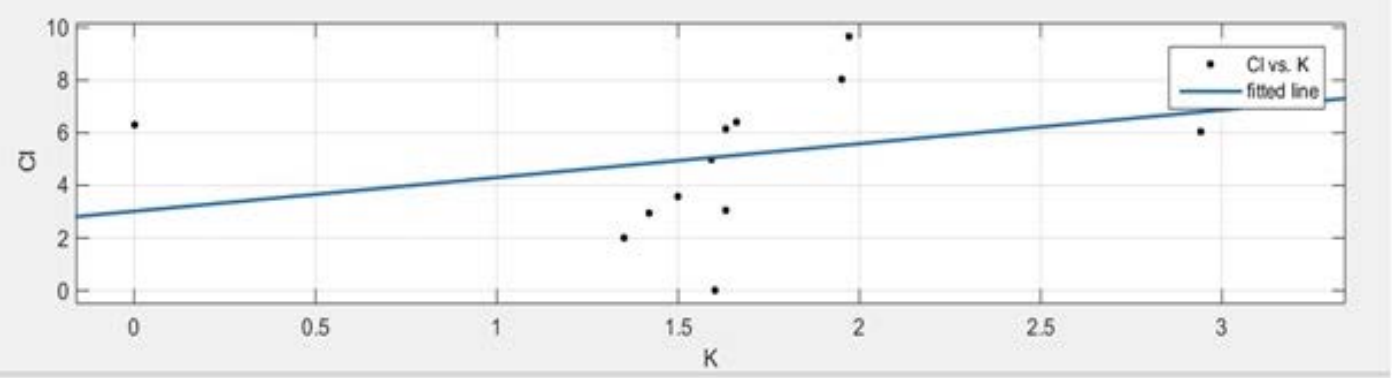

Fig 5. Linear regression plot between Cl and K+ parameters.

linear regression equations presented below (Equations 7-9) (Ortas et al., 2019).

$$a = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sum_i (x_i - \bar{x})^2} \quad (7)$$

$$\beta = \bar{y} - a\bar{x} \quad (8)$$

$$\hat{y}_i = ax_i + \beta \quad (9)$$

α = *slope*
β = *intercept*
x_i = *value of the independent variable (parameter x)*
$\bar{x}$ = *mean of the independent variable (parameter x)*
y_i = *value of the independent variable (parameter y)*
$\bar{y}$ = *mean of the dependent variable (parameter y)*
$\hat{y}_i$ = *predicted value of y for* x_i

Numerous methods have been developed by researchers for monitoring water quality. While some approaches are based on modelling systems, others include artificial intelligence-based techniques such as Artificial Neural Networks (ANN) and Genetic Algorithms (GA) (Yakut & Baran, 2025). In addition, statistical and analytical methodologies are used to examine the correlations between water quality measurements. In addition, statistical and analytical methods are widely used to examine correlations among water quality parameters and to analyse their interrelationships. As a result of these studies, prediction models have been developed that enable monitoring of water quality based on one or more parameters (Burchard-Levine et al., 2014; Sakizadeh, 2016; Barcellos & Souza, 2022).

Figure 5 presents the residual plot illustrating the linear relationship between Cl^- and K^+. The described procedure was systematically applied to all selected water quality parameters, including outlier detection and removal, missing value estimation, and data visualization. Throughout this process, missing values were imputed, resulting in a reorganized and refined dataset. The statistical analysis results of the processed dataset are summarized in Table 5, providing an overview of the data integrity and consistency achieved through these preprocessing steps.

According to the table, the variance of the Fe parameter was calculated as 143,283.4189 and the coefficient of variation (CV) was relatively high at 0.6234, indicating that the Fe data exhibit a wide distribution around the mean. For the NTU parameter, the CV was 1.215 and the variance was 6,940.662. These high variance and CV values are related to the substantial seasonal fluctuations in suspended solids in the river. Contributing factors include the study area's abundant rainfall, dense forest cover, and flood events that transport large amounts of material into the stream. Consequently, significant variations in turbidity values are expected . To evaluate the reliability of the rearranged dataset, statistical comparisons of these parameters were performed with samples obtained from other monitoring stations along the river . The analyses confirmed that the parameters in question were consistent with data from other stations.

Table 5 The statistical table of the final version of the data set, purified from outliers and imputed values for missing parameters.

unit	Parameters	Mean	max	min	Difference %	Var	CV
mg/L	Cl-	5.6010	9.6700	2.0300	376.3547	5.0113	0.3997
µg/L	Fe	607.2180	1408.0000	151.1000	831.8332	143,283.4189	0.6234
mg/L	K+	1.6632	1.9700	1.3500	45.9259	0.0335	0.1100
mg/L	Na+	13.4800	22.0697	6.5100	239.0130	20.5982	0.3367
mg/L	SO4=	37.8713	65.0000	17.5000	271.4286	200.9421	0.3743
mg/L	TKN	0.2842	0,6400	0.1000	540.0000	0.0312	0.6212
mg/L	TN	0.6538	1.0500	0.4600	128.2609	0.0334	0.2797
NTU	Turb	68.5561	236.0000	0.8000	29400	6940.6624	1.2152

Table 6: MSE values according to parameters over their normalized values

Parameters	Cl-	Fe	K+	Na+	SO4=	TKN	TN	Turb
MSE	0.01036	0.09789	0.13518	0.08733	0.01968	0.01381	0.06179	2.04E-05

Discussion

Water quality parameters play a critical role in monitoring river pollution. These parameters need to be analyzed regularly to track changes and obtain accurate information on the river's condition. However, collecting and analyzing these data is both time-consuming and costly. In addition to the expenses associated with analysis materials, the need for skilled personnel represents a significant financial burden.

The main problems encountered in this process include: In on-site analyses, some parameter values may be recorded incorrectly due to seasonal variations.

Seasonal effects and prevent proper storage of samples for certain parameters, leading to analysis failures and incomplete datasets.

In this study, various solutions were explored to address the three main problems outlined above, and the most suitable approaches were tested using actual data. To minimize analytical errors, the widely used boxplot method was applied to detect and remove potential outliers from the dataset.

Although various methods exist in the literature for handling missing data, linear regression analysis was considered the most suitable approach for data exhibiting seasonal fluctuations is. During the imputation process, correlation analysis was first performed to identify relationships between parameters. Linear regression models were then created using the parameters with high correlations, and missing values were estimated using the corresponding regression equations. This procedure was repeated for each missing data point, ultimately eliminating all deficiencies in the dataset. However, a significant portion of the data for February 2013 was missing. Because imputation using related parameters was not possible for this month, all associated observations were removed from the dataset. All statistical analyses, calculations, and visualizations in this study were performed using **Minitab**, **Excel**, and **MATLAB**.

The Mean Squared Error (MSE) values calculated for Cl^-, Fe, K^+, Na^+, SO_4^{2-}, TKN, TN, and Turbidity quantify the average squared differences between estimated and observed values, providing a measure of the model's predictive accuracy for each parameter (Nicolson & Paliwal, 2019). As presented in Table 6, the evaluation of both raw and preprocessed datasets revealed that Turbidity (Turb) exhibited the lowest MSE value (2.04E-05), indicating highly accurate estimations. In contrast, K^+ displayed the highest MSE value (0.13518), reflecting relatively greater deviation from the observed data compared to the other parameters.

Due to differences in measurement units among the selected parameters, noticeable variations were observed in their numerical scales. To facilitate comparison of parameter adjustments, all values were normalized, and MSE values were subsequently evaluated for both datasets. This normalization process eliminated interpretation difficulties arising from unit discrepancies (e.g. Fe in µg/L, versus K^+ in mg/L) and enabled all parameters to be evaluated on a standardized scale.

In this study, after identifying outliers and missing values within the dataset, the relationships among parameters were examined, and regression models were developed for each parameter based on those with high correlations. The datasets were then refined through imputation using the derived regression models, followed by statistical comparisons to assess the extent of the impact on each parameter. To ensure reproducibility and facilitate interpretation, the procedures were repeated using normalized data. The results indicated that the adjustments performed on the dataset were highly consistent. Final evaluations are summarized in Tables 5 and 6.

Recommendation

In many countries, routine monitoring of water resources is governed by legal regulations and implemented by authorized institutions. To enhance water quality monitoring , it is recommended to adopt more comprehensive and innovative approaches. In particular, artificial neural networks (ANN), genetic algorithms, and other AI-based methods have significant potential to improve existing water quality monitoring processes. The integration of these methods can contribute to a more accurate and rapid assessment of the status of water resources. For broader implementation, future guidance from regulatory bodies and increased investments in infrastructure will be essential.

This study focused on data from a single monitoring station of the Araç River, based on analyses performed by a legal water monitoring organization. In the first phase, common issues such as data pollution and outliers were identified using the box plot method, and these values were removed to obtain a more consistent dataset. Missing data resulting from seasonal effects and extreme values were subsequently imputed, producing a reliable and clean dataset. This approach emphasizes the importance of data quality in water resources management and monitoring processes and provides a foundation for more robust analyses.

In this study, eight key parameters - chloride ion (Cl-), iron (Fe), potassium ion (K+), sodium ion (Na+), sulphate ion (SO_4^{2-}) total Kjeldahl nitrogen (TKN), total nitrogen (TN) and turbidity (Turb) were selected from the organized dataset. Given the critical role of these parameters in water quality assessment, future studies are encouraged to include additional water quality indicators. Moreover, the use of datasets covering longer time periods in monitoring processes could enable more detailed analysis of seasonal and long-term trends. Employing multidisciplinary approaches and advanced modeling techniques for the protection and sustainable management of water resources can increase the effectiveness of monitoring processes and provide more comprehensive information to policymakers.

To develop more sensitive and reliable prediction models, it is of great importance that the datasets used in the analyses have broader spatial coverage. Incorporating samples from additional monitoring stations can enable models to capture regional variations more accurately. Spatially diverse datasets

allow predictive models to learn complex relationships and patterns, thereby enhancing performance. Furthermore, larger and more comprehensive datasets improve the generalization capacity and accuracy of predictive models, thus contributing to better adaptation of models to varying environmental conditions. Consequently, it is recommended that water quality monitoring studies expand data collection and integrate them with existing methods.

Statements & Declarations

The author declares no conflict of interest. This study does not contain any private or sensitive information, as all data were anonymized and presented in tabular form. Additionally, artificial intelligence (AI) was not used for analyses and methodological procedures.

Declaration of generative AI and AI-enabled technologies in the writing process

During the preparation of this study, the author used artificial intelligence tools to assist with the text editing, particularly paragraph layout and structure. The author reviewed and revised the content and assumes full responsibility for the accuracy and integrity of the publication.

Ethical Approval

This research, entitled „Preparation of Water Quality Parameters Before Analysis," involved a thorough examination of dataset preparation steps and experimental applications using real data. The data were obtained from the General Directorate of State Hydraulic Works. Statistical properties of the data are reported to ensure study integrity, and no personal information was included.

Consent to Participate

No personal data was used in the study.

Consent to Publish

The author affirms that the manuscript has not been submitted to any preprint server prior to this submission.

Authors Contributions

This study was conducted by a single researcher.

Funding

The author declares no financial interests related to this study.

Competing Interests

The author declares no financial support or competing interests related to this study.

Data Availability Statement

The dataset used in this study comprises water quality parameters collected from a selected river system. Due to institutional and data privacy policies, the raw dataset is not publicly available. However, all analyses were conducted using statistically summarized forms of the data, such as mean, standard deviation, and range values. These summaries are sufficient to replicate the preprocessing procedures (e.g., outlier detection, imputation, and normalization) and to evaluate model performance.

There are no restrictions preventing access to the summarized data, and no issues affecting the validity or reproducibility of the findings presented in this article.

Reference

Addico, G., Hardege, J., Komarek, J., Babica, P. & de Graft-Johnson, K. (2006). Cyanobacteria species identified in the Weija and Kpong reservoirs, Ghana, and their implications for drinking water quality with respect to microcystin. *African Journal of Marine Science*, *28*(2), 451–456. https://doi.org/10.2989/18142320609504196

Ahmad, S., Khan, I. H. & Parida, B. P. (2001). Performance of Stochastic Approaches for Forecasting River Water Quality. *Wat. Res*, *35*(18), 4261–4266. https://doi.org/10.1016/S0043-1354(01)00167-1

Azhar, S. C., Aris, A. Z., Yusoff, M. K., Ramli, M. F. & Juahir, H. (2015). Classification of River Water Quality Using Multivariate Analysis. *Procedia Environmental Sciences*, *30*, 79–84. https://doi.org/10.1016/j.proenv.2015.10.014

Barcellos, D. da S. & Souza, F. T. de. (2022). Optimization of water quality monitoring programs by data mining. *Water Research*, *221*. https://doi.org/10.1016/j.watres.2022.118805

Betrie, G. D., Sadiq, R., Tesfamariam, S. & Morin, K. A. (2016). On the Issue of Incomplete and Missing Water-Quality Data in Mine Site Databases: Comparing Three Imputation Methods. *Mine Water and the Environment*, *35*(1), 3–9. https://doi.org/10.1007/s10230-014-0322-4

Burchard-Levine, A., Liu, S., Vince, F., Li, M. & Ostfeld, A. (2014). A hybrid evolutionary data driven model for river water quality early warning. *Journal of Environmental Management*, *143*, 8–16. https://doi.org/10.1016/j.jenvman.2014.04.017

Chen, X., Strokal, M., van Vliet, M. T. H., Fu, X., Wang, M., Ma, L. & Kroeze, C. (2022). In-stream surface water quality in China: A spatially explicit modelling approach for nutrients. *Journal of Cleaner Production*, *334*(May 2021), 130208. https://doi.org/10.1016/j.jclepro.2021.130208

Dawson, R. (2011). How significant is a boxplot outlier? *Journal of Statistics Education*, *19*(2). https://doi.org/10.1080/10691898.2011.11889610

Dong, Y. & Peng, C. Y. J. (2013). Principled missing data methods for researchers (Expectation Maximaization explained). *SpringerPlus*, *2*(1), 1–17.

Efstathiou, C. E. (2006). Estimation of type I error probability from experimental Dixon's "Q" parameter on testing for outliers within small size data sets. *Talanta*, *69*(5), 1068–1071. https://doi.org/10.1016/j.talanta.2005.12.031

Fouad, K. M., Ismail, M. M., Azar, A. T. & Arafa, M. M. (2021). Advanced methods for missing values imputation based on similarity learning. *PeerJ Computer Science*, *7*, 1–38. https://doi.org/10.7717/PEERJ-CS.619

Garlits, J., McAfee, S., Taylor, J.-A., Shum, E., Yang, Q., Nunez, E., Kameron, K., Fenech, K., Rodriguez, J., Torri, A., Chen, J., Sumner, G. & Partridge, M. A. (2023). Statistical approaches for establishing appropriate immunogenicity assay cut points: Impact of sample distribution, sample size, and outlier removal. *The AAPS Journal, 25*(37). https://doi.org/10.1208/s12248-023-00806-5

Guo, Y.-H., Fan, X.-Y., Zhang, L., Fan, H. & Xu, Y.-J. (2015). Determination of La/CeO_2 content in ilmenite electrode coating. *Rare Metals, 34*(7), 505–509. https://doi.org/10.1007/s12598-014-0406-0

Holt, B. & Benfer, R. A. (2000). Estimating missing data: An iterative regression approach. *Journal of Human Evolution*, *39*(3), 289–296. https://doi.org/10.1006/jhev.2000.0418

Horvat, Z., Horvat, M., Pastor, K., Bursić, V. & Puvača, N. (2021). Multivariate analysis of water quality measurements on the danube river. *Water (Switzerland)*, *13*(24), 1–20. https://doi.org/10.3390/w13243634

Isaac, R. & Siddiqui, S. (2022). Application of water quality index and multivariate statistical techniques for assessment of water quality around Yamuna River in Agra Region, Uttar Pradesh, India. *Water Supply*, *22*(3), 3399–3418. https://doi.org/10.2166/WS.2021.395

Islam Khan, M. S., Islam, N., Uddin, J., Islam, S. & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003

Jancosek, M. & Pajdla, T. (2014). Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces. *International Scholarly Research Notices, 2014*, Article ID 798595, 20 pages. https://doi.org/10.1155/2014/798595

Kadengye, D. T., Cools, W., Ceulemans, E. & van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior Research Methods*, *44*(2), 516–531. https://doi.org/10.3758/s13428-011-0157-x

Li, X., Ding, J. & Ilyas, N. (2021). Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. *Water Science and Technology: Water Supply*, *21*(3), 1291–1312. https://doi.org/10.2166/ws.2020.381

Liu, C. (2001). A comparison of five distance-based methods for spatial pattern analysis. *Journal of Vegetation Science*, *12*(3), 411–416. https://doi.org/10.2307/3236855

Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A. & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems*, *42*(5), 1–17. https://doi.org/10.1007/s10916-018-0940-7

Nicolson, A. & Paliwal, K. K. (2019). Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, 111, 44–55. https://doi.org/10.1016/j.specom.2019.06.002

Ortas, E., Burritt, R. L. & Christ, K. L. (2019). The influence of macro factors on corporate water management: A multi-country quantile regression approach. *Journal of Cleaner Production*, *226*, 1013–1021. https://doi.org/10.1016/j.jclepro.2019.04.165

Roger A. Falconer (1992). Flow and water quality modelling in coastal and inland water. *Journal of Hydraulic Research,* (30) issue 4, page: 437-452. https://doi.org/10.1080/00221689209498893

Sadiq, Q., Ezeamaka, C. K., Daful, M. G. & Mustafa, I. A. (2022). Evaluation of the Water Quality of River Kaduna, Nigeria Using Water Quality Index. *Environmental Technology and Science Journal*, *13*(1), 28–40. https://doi.org/10.4314/etsj.v13i1.3

Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, *2*(1). https://doi.org/10.1007/s40808-015-0063-9

Salgado, C. M., Azevedo, C., Proen, H. & Vieira, S. M. (2016). Secondary Analysis of Electronic Health Records. *Secondary Analysis of Electronic Health Records*, 1–427. https://doi.org/10.1007/978-3-319-43742-2

Tripathi, M. & Singal, S. K. (2019). Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecological Indicators*, *96*, 430–436. https://doi.org/10.1016/j.ecolind.2018.09.025

Wang, J., Zhang, Y., Cao, H. & Zhu, W. (2012). Dimension reduction method of independent component analysis for process monitoring based on minimum mean square error. *Journal of Process Control*, *22*(2), 477–487. https://doi.org/10.1016/j.jprocont.2011.11.005

Wilrich, P. T. (2013). Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic. *AStA Advances in Statistical Analysis*, *97*(1), 1–10. https://doi.org/10.1007/s10182-011-0185-y

Wu, Y., Kihara, K., Takeda, Y., Sato, T., Akamatsu, M. & Kitazaki, S. (2020). The relationship between drowsiness level and takeover performance in automated driving. In H. Krömker (Ed.), *Human interaction and emerging technologies: Proceedings of the 2nd International Conference on Human Interaction and Emerging Technologies (IHIET 2020)* (Lecture Notes in Computer Science, Vol. 12213, pp. 125–142). Springer. https://doi.org/10.1007/978-3-030-50537-0_11

Xie, W., Chkrebtii, O. & Kurtek, S. (2020). Visualization and Outlier Detection for Multivariate Elastic Curve Data. *IEEE Transactions on Visualization and Computer Graphics*, *26*(11), 3353–3364. https://doi.org/10.1109/TVCG.2019.2921541

Yakut M.Ş & Baran B. (2025). A Hybrid Machine Learning and Stochastic Modeling Framework for Probabilistic Reliability Analysis of Kızılırmak River Water Quality. *Water Environment Research*, Vol 97 Issue 9 Sep 2025, 117536. https://doi.org/10.1002/wer.70169

Yang, D., Luan, W., Li, Y., Zhang, Z. & Tian, C. (2023). Multi-scenario simulation of land use and land cover based on shared socioeconomic pathways: The case of coastal special economic zones in China. *Journal of Environmental Management*, *335*(January), 117536. https://doi.org/10.1016/j.jenvman.2023.117536

Yang, R. (2022). *Analyses of Approaches to Deal with Missing Data in Water Quality Data Set*. https://doi.org/https://doi.org/10.2991/aebmr.k.220405.184

Zhang, Y. & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, *128*, 63–72. https://doi.org/10.1016/j.future.2021.09.033