Applications of Missing Data Imputation Methods in Wastewater Treatment Plants

A Systematic literature Review using the Kitchenham Method

Abdellah Chaoui[1], Kaoutar Rebija[2], Kaoutar Chkaiti[3], Mohammed Laaouan[4],
Rqia Bourziza[5], Karima Sebari[6], Wafae Elkhoumsi[7]
IAV Hassan II, Rabat, Morocco[1, 2, 3, 5, 6, 7]
International Institute for Water and Sanitation, Rabat, Morocco[4]

***Abstract*—Missing data pose a big challenge in the field of wastewater treatment, representing a frequent issue in data quality that can result in misleading analyses and compromised decision-making accuracy. The initial step in data preprocessing involves the estimation and handling of missing values. The primary aim to conduct a comprehensive examination of the existing research concerning missing value imputation in wastewater treatment plants (WWTPs). The focus is specifically on identifying and outlining various imputation techniques employed in this field, while paying close attention to their respective strengths and limitations. To ensure a methodical approach, this study adopts the systematic literature review (SLR) using Kitchenham's guidelines. In order to gather relevant and up-to-date papers, the research leverages the scientific database "Scopus" to retrieve and analyze all pertinent papers during the search process. By doing so, this research aims to contribute valuable insights into the different strategies used for imputing missing values in WWTPs and to shed light on their practical implications and potential drawbacks. Form 599, a total of 16 research papers were selected to assess the review questions. Finally, several recommendations were given to address the limitations identified in the reviewed studies and to contribute to more accurate and reliable data analysis and decision-making in the wastewater treatment domain.**

***Keywords*—Systematic literature review; kitchenham' method; wastewater treatment; imputation methods; missing data**

I. Introduction

The presence of missing values presents a notable obstacle to ensuring the data quality of WWTPs datasets. Despite the presence of well-designed data collection systems in many treatment plants, the attention given to data quality is often inadequate [1].

This problem affects the performance of data analytics, leading to increased bias and decreased accuracy. The presence of missing values can be attributed to various factors, such as events causing measurement failures, holidays, and shifts with less experienced personnel [2]. As a result, gaps or discontinuities arise in the data records, severely hindering the modeling and identification of the process [3].

According to Rubin's [4], missing values can be classified into three main mechanisms, presented below:

- "Missing completely at random" (MCAR): The absence of data has no correlation on the missing data itself.
- "Missing at random" (MAR): Missing values that exhibit a relationship with the observed values.
- "Missing not at random" (MNAR): It applies when neither of the previous mechanisms is valid, and the missing values are typically associated with unobserved predictors or the missing value itself [5].

Over the past, researchers have shown considerable interest in that type of problem. Common approaches for handling missing data include deletion and imputation. Imputing missing data is a crucial step as any data analysis with incomplete datasets would yield invalid conclusions. Ignoring this step can result in biased estimations [6] - [7] While missing data imputation is a well-established technique in data analysis, its application in the context of WWTPs remains relatively unexplored. Existing literature on WWTPs often overlooks or inadequately addresses the critical issue of missing data imputation. Consequently, there is an urgent need to systematically review and evaluate the available missing data imputation methods specifically for WWTP datasets.

There exists an urgent requirement for a dependable and efficient approach to substitute missing data, as this is crucial for accurately assessing the variability of plant influent data. By doing so, more precise design proposals and performance evaluation reports can be generated; leading to improved decision-making processes in wastewater treatment plants [2].

This paper aims to analyse the techniques employed for imputing missing data in WWTPs and review the available methods through imputation. The following sections will present a concise overview of the relevant literature and the quantity of research studies that concentrate on imputing missing data in WWTPs.

In this paper, Section I offers an introduction, Section II outlines the research methodology, and Section III discusses the research findings. A discussion of the results is provided in Section IV, while Section V concludes the article.

II. Research and Review Method

In this section, the methodology used is this paper is Kitchenham' method as presented below:

A. Planning the Review

In this section, the review methodology essential for conducting the systematic literature review (SLR). This entails formulating research questions aligned with the primary objective of the review, devising a robust search strategy, and crafting a comprehensive review protocol. Each of these aspects plays a critical role in ensuring the rigor and effectiveness of the SLR process.

1) Research questions: The primary objective of this review is to examine the current literature concerning imputation methods used in the wastewater field. The specific research questions (RQs) are outlined below:

a) What are the existing methods applied in wastewater treatment plants?

b) How effective are the existing methods in handling missing data challenges?

c) What are the limitations of those techniques applied specifically in the context of wastewater treatment plants?

These RQs serve as the guiding framework for this study and facilitate a systematic and thorough analysis of the relevant literature. RQ1 identifies and documents the various techniques used in the context of wastewater treatment plants to handle and optimize missing data imputation. RQ2 focuses on assessing the effectiveness of these existing methods in effectively addressing the challenges posed by missing data in wastewater treatment plants. Lastly, RQ3 aims to analyze the limitations of different missing data imputation techniques specifically within the context of wastewater treatment plants.

2) Search strategy: The Scopus database was chosen because of the wide range of academic literature from a variety of fields, including engineering and environmental sciences. It provides a huge selection of peer-reviewed journals, conference papers, and other pertinent literature. The search string used to retrieve articles from the scientific database is described as follows: ("Imputation" OR "missing value*") AND ("Wastewater" OR "WWTP" OR "WATER").

3) Inclusion and exclusion criteria: The inclusion and exclusion criteria are illustrated in Table I. By applying these criteria, the aim was to identify and focus on the most pertinent studies that align with the research objectives and ensure the inclusion of high-quality and relevant sources in the final analysis.

4) Quality criteria: The primary objective of this section is to check that primary studies contain adequate information to address the research questions. Each criterion is labeled as 'QAC,' which stands for Quality Assessment Criteria. These criteria serve as a means to assess the quality and relevance of each primary study, ensuring that they provide sufficient insights to effectively answer the research questions.

TABLE I. The Inclusion and Exclusion Criteria.

Inclusion criteria	Exclusion criteria
Papers that address missing data imputation methods applied in the context of wastewater treatment plants.	Papers that don't address missing data imputation methods applied in the context of wastewater treatment plants
Papers published in any journal or conference proceedings and in any language.	Multiple versions of the same study and duplicate publications
Articles that are available in full text	Articles that are not available in full text

These QACs play a vital role in assessing the quality of the selected articles, enabling the determination of their overall quality and suitability for the systematic review.

The quality of studies is assessed through the following evaluation questions:

- **QAC.1** Does the paper use missing data imputation methods for the wastewater domain?
- **QAC.2** Were the key parameters containing missing values mentioned clearly in the paper?
- **QAC.3** Did the researchers explain the performance measurements used?
- **QAC.4** Does the paper cover limitations of the proposed method?

Apart from evaluating the inclusion and exclusion criteria, a thorough examination of each primary study was conducted, employing specific QAC questions [8]. In the evaluation process, each primary study was assigned a score between 0 and 1. A score of 1 indicated that the study fully addressed the QAC question, while a score of 0.5 denoted a partial answer. On the other hand, if the study failed to address the QAC question, a score of 0 was given. The cumulative score for each study was then calculated by summing the scores for all the QAC questions.

Upon completing the quality assessment for each primary study, it was observed that the total score of the selected studies exceeded 50% for each QAC, as presented in Table II.

TABLE II. Quality Assessment Criteria and Results of Selected Articles

Criteria	Responding score	Total score
QAC01	{0, 0.5, 1} (No, Partially, Yes)	16 studies (100%)
QAC02	{0, 0.5, 1} (No, Partially, Yes)	12 studies (75%)
QAC03	{0, 0.5, 1} (No, Partially, Yes)	13 studies (81.25%)
QAC04	{0, 0.5, 1} (No, Partially, Yes)	9 studies (56%)

This finding suggests that the primary studies included in the review contain substantial and relevant information.

B. Conducting the Review

The search was conducted on April 1st, 2023, without imposing any date or language restrictions. The article selection process involved applying the specified search string, resulting in 599 initial papers. From these papers, relevant information was exported to a spreadsheet, then, the search results were filtered by removing the duplicates articles and those with no abstract available, which left 593 papers. The remaining papers were subjected to a manual review of their titles and abstracts, leading to a selection of 61 papers retrieved from the water quality field, surrounding wastewater treatment. The subsequent step involved referring to and reading the full-text articles in a meticulous manner. During this process, five articles were excluded due to the unavailability of their full text. Then, after analyzing all the papers and organizing the evidence specifically related to wastewater treatment, it was determined that only 16 articles were deemed relevant and shortlisted. As a result, 16 articles successfully met the research questions and satisfied all the inclusion and quality assessment criteria outlined. The paper selection procedures are succinctly summarized in Fig. 1.

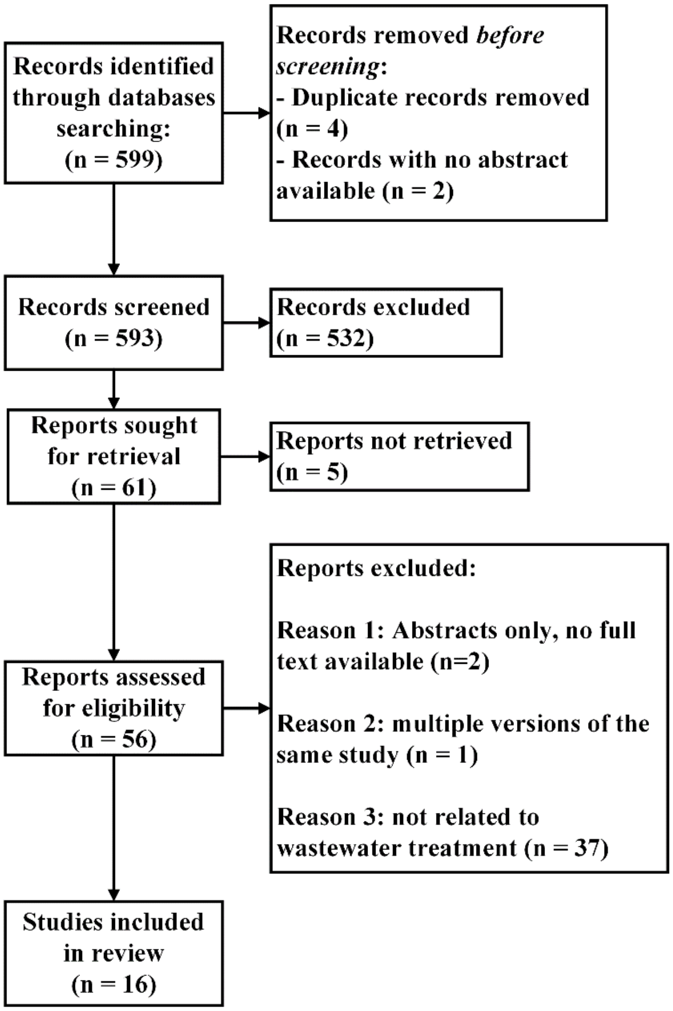
Records identified through databases searching: (n = 599)
Records removed *before screening*:
- Duplicate records removed (n = 4)
- Records with no abstract available (n = 2)
Records screened (n = 593)
Records excluded (n = 532)
Reports sought for retrieval (n = 61)
Reports not retrieved (n = 5)
Reports assessed for eligibility (n = 56)
Reports excluded:
Reason 1: Abstracts only, no full text available (n=2)
Reason 2: multiple versions of the same study (n = 1)
Reason 3: not related to wastewater treatment (n = 37)
Studies included in review (n = 16)

Fig. 1. The article selection process.

III. Findings

This section presents and analyzes the results obtained from the literature review. Findings are divided into three subsections, with the first one showcasing the various methods used. The second subsection discusses the effectiveness of these existing methods. Finally, the third subsection explores the strengths and limitations of the different methods.

A. The Identified Methods

This subsection primarily focuses on RQ1, which aims to identify the existing methods. A concise overview of the techniques employed to handle missing values in WWTPs is presented in this subsection. Fig. 2 illustrates the publication trend over time in this specific research area.

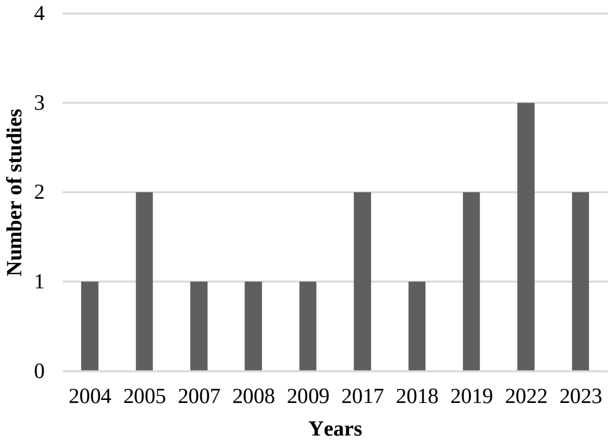

Fig. 2. Number of studies published per year.

In their study, Huo et al. [9] the Two-directional Exponential Smoothing (TES) method to impute missing data in a WWTP. In another study in 2010, they applied also the TES and TESWN (Two-directional Exponential Smoothing with Nearest Observations). The TES method involves generating a hypothetical complete data set using the average of nearest observations (ANO) method. It then forecasts the missing values using an exponential smoothing algorithm in both the forward direction (Forward ES) and the backward direction (Backward ES). The missing data points' ultimate replacement values are obtained by averaging the estimates derived from both the forward and backward exponential smoothing forecasts. The TESWN method shares similarities with the TES approach but includes a white noise term to handle random effects observed in the data, which might not be adequately addressed by the autocorrelation function [10].

Zhang et al. [11] choose the self-organizing map (SOM) model to impute the missing data by training the model using available data and then presenting the depleted vector to the SOM to identify its best matching unit (BMU). The missing variable values are acquired by referring to their respective values in the BMU.

In the study conducted by Villez et al. [12], the imputation of missing values was carried out through a backward calculation based on scores obtained from the inverse Principal Component Analysis (PCA). The scores were estimated using the single component-projection method [13]. Negative estimates for concentrations were rectified by adjusting them to zero. After this adjustment, any remaining missing variables were subsequently re-estimated using the same methodology.

Borzooei et al. [14] investigates the application of the Cubic Hermite interpolation method for filling in missing values within data. This method is particularly suitable for datasets characterized by rapid and non-linear changes. It employs a third-degree polynomial function to approximate the missing value based on the surrounding data points. To use this method effectively, the data must exhibit continuity, and the function must be differentiable over the relevant interval.

Furthermore, a minimum requirement of having at least two adjacent points to the missing value is necessary for performing the interpolation.

De Mulder et al. [15] employed various filling algorithms to address missing data gaps. These algorithms included interpolation, the utilization of daily average values, and the incorporation of values from the previous day, correlation-based approaches, and the application of the influent model.

Azizoğlu et al. [16] conducted a study, employing six distinct machine learning algorithms, including Linear Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), and Adaboost, to estimate missing pH data in two distinct datasets.

Phung et al. [17] utilized the multiple imputation procedure to handle missing values. Instead of assigning a single value to each missing data point, they employed Rubin's [4] multiple imputation approach, which replaces each missing value with a set of plausible values. Afterwards, these imputed datasets were analyzed using standard procedures designed for complete data, and the results obtained from these analyses were combined. The process of combining results from different imputed datasets remained consistent, regardless of the specific complete-data analysis method used.

Pascual-Pañach et al. [18] explores the utilization of the Case-Based Reasoning (CBR) approach for online imputation of missing values. This approach addresses the data imputation problem by leveraging past solutions to analogous problems. By employing the Case-Based Reasoning (CBR) principle in data imputation, values from similar historical scenarios are leveraged to replace incorrect or missing values. To improve the effectiveness of the data imputation process, optimal case feature weights are determined using genetic algorithms (GA). The proposed methodology is validated using real data obtained from an operational WWTP process.

Xiao et al. [19] suggested the implementation of Auto-associative Neural Networks (ANN) along with a recursive minimization strategy to address missing values in fault diagnosis for wastewater processes. The ANN is trained using available data to capture the inherent patterns and interrelationships among the variables. Meanwhile, the recursive minimization strategy is employed to iteratively update the missing values based on the learned patterns until convergence is achieved.

Ba-Alawi et al. [20] employed R2AU-Net, an automated data reconciliation and imputation approach tailored to handle missing and faulty data in the membrane bioreactor (MBR) process. This method used a recurrent residual convolutional neural network (CNN) with an attention gate (AG) connection to effectively impute consecutive missing data and reconcile faulty sensors in the MBR process. During training, the model employs backpropagation with the Adam optimizer and mean squared error (MSE) as the activation function. The input and output sizes are set to 18, corresponding to the number of studied sensors. To enhance performance and training speed, the R2AU-Net model incorporates batch normalization (BN) in each R2CL block. The training process is carried out using the Keras library and TensorFlow backend, with 370 epochs and a batch size of 24.

In their research article, Han et al. [21] utilized the univariate imputation method (UIM) in conjunction with the SSD method and SVR model. The UIM decomposes the time series into seasonal, trend, and remainder components and employs specific imputation methods for each component. The SSD method addresses missing values in the seasonal component by identifying repeating patterns. On the other hand, the SVR model is responsible for imputing missing values in the trend and remainder components. By integrating the imputation results, UIM effectively handles missing values in WWTP time series data.

Safford et al. [22] focuses on the application of the EM-MCMC algorithm for estimating missing values in wastewater qPCR data. The proposed method involves a systematic process that begins with the initialization of hyperparameters. Subsequently, Monte Carlo samples of the latent parameters are generated using the Markov Chain Monte Carlo (MCMC) technique. Finally, the maximum likelihood estimates of the hyperparameters are computed based on the sampled latent parameters.

Yan et al. [23] used the Non-Linear Decreasing Inertia Weight Particle Swarm algorithm (NLDIWPSO) based optimal Support Vector Regression (SVR) approach to impute missing values. For abnormal values and missing values with a non-continuous distribution over time, they used the average of non-abnormal data for a period of three days before and after to fill the gaps. Conversely, for abnormal values and missing values with a continuous distribution over time, the NLDIWPSO-based optimal SVR method was employed for forecasting purposes.

Oliveira-Esquerre et al. [24] used linear interpolation for estimating missing values, citing its simplicity [24]. This method was applied with the constraint that it was only employed for a maximum of two consecutive missing values.

In summary, various methodologies have been employed to handle missing data in WWTP studies and there is no imputation technique consistently outperforms every other. The Table III below summaries the missing data imputation methods with their used respective key parameter in a WWTP.

TABLE III. MISSING DATA IMPUTATION METHODS, KEY PARAMETERS, AND RELEVANT STUDIES

Missing Data imputation method	Parameter	Studies
R2AU-Net	- pH, - Dissolved oxygen (DO)	[20]
Univariate Imputation Method (UIM) (SSD method & SVR model)	- pH, - SS - BOD - NH4	[21]
Linear regression (LR) K-Nearest Neighbors (KNN) Random Forest (RF) Adaboost Decision Tree (DT) Support Vector Machines (SVM)	- pH	[16]
Coupling the Expectation Maximization (EM) algorithm with the Markov Chain Monte Carlo (MCMC) method (EM-MCMC)	- N1 - N2 - PMMoV concentrations	[22]
Case-Based Reasoning (CBR) approach	real data from a WWTP	[18]
Cubic Hermite interpolation method	- COD - N-NH4 - TSS - T° - Influent flow rate (Qin)	[14]
-For abnormal values and missing values that are discontinuously distributed over time: the average of the non-abnormal data for three days before and after was used to fill it. -For abnormal values and missing values that are continuously distributed over time: the Non-Linear decreasing inertia weight particle swarm algorithm (NLDIW-PSO) based optimal Support Vector Regression (SVR) was used to forecast.	- PH - COD - BOD5 - TP - TN - NH4-N	[23]
Interpolation Correlation Daily average Day before Influent model	- CODt - NH4	[15]
Auto-associative neural network (ANN) with a recursive minimization strategy		[19]
Multiple imputation (MI)	- Atenonol - Codeine - Cafeine - Hydrochlorthiazide - Acesulfame - Salicylic Acid - Carbamazepine - Naproxen	[17]
SOM	- Ammonia-nitrogen, - Soluble reactive phosphorus (SRP), - SS, - COD - BOD5	[11]
TES and TESWN	- TSS	[10]
Kohonen self-organizing map (KSOM), unsupervised neural networks	- flow rate - Influent BOD, and TSS - WAS rate - mixed liquor suspended solids MLSS - return activated sludge mixed liquor suspended solids - stirred sludge volume index SSVI - sludge age - food to microorganism's ratio F/M - effluent flow, BOD, COD and TSS.	[3]
PCA projection method	- nitrogen species	[12]
Two-directional exponential smoothing (TES)	- BOD5 - TSS - NH4-N	[9]

Linear interpolation	- BOD - COD - COL color - COND conductivity - NAM inlet ammonia concentration - NN inlet nitrate concentration - pH - PAP paper production - PULP pulp production - RF rainfall - T° - TSS	[24]

B. The Challenges

Existing methods have varying degrees of effectiveness in handling missing data challenges. Ba-Alawi et al. [20] found that the R2AU-Net model exhibited the highest imputation performance for missing data, even when the missing interval increased to 50%. It outperformed conventional methods like PCA and DPCA, as well as neural methods like AE and VAE, with the lowest Mean Absolute Error (MAE) value of 0.31 mg/L. Consequently, the R2AU-Net missing data imputation approach is regarded as highly effective in tackling missing data issues. Additionally, the paper explores the use of the PCA projection method to estimate missing data in the SHARON process. Villez et al. [12] showed that the estimation of missing data related to nitrogen species enhances the performance of a dynamic PCA model. However, despite this improvement, the impact of data gaps remains significant, as the undetected failure ratio nearly doubles when no estimates are employed.

Therefore, while PCA can be helpful in handling missing data challenges, it may not completely solve the problem.

According to the experimental results presented by Han et al. [21], the UIM method, as proposed in the study, proves to be effective in imputing missing data in WWTP time series compared to the other seven competitors examined. In the testing phase, UIM underwent evaluation using six distinct WWTP time series, and the outcomes demonstrated that it strikes a well-balanced trade-off between imputation accuracy and processing time. Notably, UIM and NA.linear exhibit remarkable performance concerning Root Mean Squared Error (RMSE) when confronted with significant intervals of missing data. Moreover, the proposed UIM exhibits the capability to handle a maximum missing ratio of 45%.

Based on Azizoğlu & Ünsal's study in 2022 [16], machine learning algorithms proved to be highly effective in predicting missing pH data. The performance was evaluated using the MAE, mean squared error (MSE), and RMSE as metrics. The results indicated that the SVM (Support Vector Machine) algorithm outperformed other algorithms in all three-performance metrics for both datasets. Consequently, the method of imputing missing data using machine learning algorithms was found to be a successful approach in addressing issues related to missing data.

Pascual-Pañach et al. [18] findings, the performance of the proposed methodology of using a Case-Based Reasoning (CBR) approach was improved by obtaining optimal case feature weights using genetic algorithms (GA). In comparison to non-calibrated CBR imputation systems, the technique was deemed highly effective, as the RMSE of the estimation using weighted features was nearly 40% lower than the non-weighted estimation when employing temporal CBR (TCBR). Moreover, the TCBR approach exhibited even better performance, with an RMSE approximately 60% lower than the calibrated CBR approach.

Yan et al. [23] assessed the performance of the NLDIW-PSO based optimal SVR machine learning model for imputing missing data using the coefficient of determination and Pearson's correlation coefficient. They found that this method achieved the highest prediction accuracy when compared to other data-driven models. Furthermore, the experimental results highlighted several advantages of the proposed model, including enhanced stability and time efficiency compared to traditional data-driven models like BP ANN, Bayesian network model, and Decision Tree model. Consequently, the NLDIW-PSO algorithm demonstrated strong performance in imputing missing data.

According to De Mulder et al. [15], the reliability of different missing data imputation methods was tested for different types of data. The results showed that using the influent model to fill gaps in the data yielded the highest reliability, while linear interpolation was also effective for smaller gaps in the data. However, all filling algorithms seem to do what they were designed for in a satisfying way, and the choice of method may depend on the specific dataset and the purpose of the analysis.

Xiao et al. [19] demonstrated the effectiveness of the current auto-associative neural network (ANN) with a recursive minimization strategy in handling missing data, as well as overcoming the Gaussian assumptions of traditional multivariate statistics models. Through simulation studies, the proposed method showcased good performance, even in scenarios with significant amounts of missing data in both the BSM1 simulation platform and real WWTP datasets.

Zhang et al. [11] concluded that the self-organizing map (SOM) model was an accurate and effective method for predicting missing values and replacing outliers in the integrated constructed wetland (ICW) data set. The SOM model demonstrated resilience to missing values and effectively processed incomplete data sets, resulting in accurate predictions. For ammonia-nitrogen, SRP, COD, SS, and BOD, the proportions of missing values and outliers were approximately 4%, 3%, 41%, 54%, and 61%, respectively. According to Rustum & Adeloye [3], the Kohonen self-organizing map (KSOM) proves to be a valuable tool for imputing missing values and handling outliers in high-dimensional datasets. The results demonstrate that the KSOM outperforms univariate prediction models based on linear regression and backpropagation ANN. Among the three approaches, the linear regression model displayed the least performance. Evaluation of the KSOM's performance using mean square error (MSE) and average absolute error (AAE) as parameters revealed that the KSOM achieved lower MSE and AAE values compared to regression and ANN. Additionally, a notable advantage of the KSOM is that the same map can be used to predict any missing value in any variable.

Huo et al. [10] showed that the TES and TESWN methods proposed in the article are effective in handling missing data challenges. The TES method is ideal when the objective is to minimize the average error linked to missing values, whereas the TESWN method is more suitable for quantifying the level of uncertainty associated with the missing values. ANO and AVE were utilized as benchmarks to compare the performance of the TES and TESWN methods. In their study, Huo et al. [9] pointed out that several commonly used methods for estimating missing values rely on the assumption of MCAR (missing completely at random), which is not applicable in their data due to the presence of a regular pattern of missing data. To address this challenge, the TES method is presented as a potential solution. The authors employed performance parameters such as R2, RMSE, and Mean Absolute Percent Error (MAPE) to assess the effectiveness of the time series models they developed.

C. Limitations

The UIM method proposed in the study can handle missing data up to a maximum ratio of 45%. However, when the missing ratio exceeds 45%, the UIM method may not generate an appropriate result [21].

According to Azizoğlu & Ünsal's [16], linear regression (LR) assumes a linear relationship between the variables and may not work well with non-linear data. K-Nearest Neighbors (KNN) is sensitive to the choice of k value and may not work well with high-dimensional data. Random Forest (RF) and Adaboost may overfit the data if the number of trees is too high. Decision Tree (DT) may suffer from overfitting and instability if the tree is too deep or complex. Support Vector Machines (SVM) can be computationally expensive for large datasets and may not work well with imbalanced data.

For Safford et al. [22], the EM-MCMC method encounters limitations in terms of incomplete comparisons due to sampling zones being added over time, and the need for further testing of the effect of different data groupings on model performance.

The Cubic Hermite interpolation method assumes that the data is smooth and continuous [14]. This technique may not be effective for datasets with a significant number of missing values. It also may not be suitable for datasets with irregular time intervals between observations. The Cubic Hermite interpolation may introduce errors when dealing with data containing outliers or extreme values [14].

Per the findings of De Mulder et al. [15], interpolation operates under the assumption of a linear relationship between missing data and the surrounding data points. Correlation-based approaches, on the other hand, rely on the presence of a correlation that may not be existent. Daily average estimation may fall short in capturing the full variability of the data, leading to potential biases. Likewise, relying on the previous day's data assumes a level of similarity that may not always be valid. Finally, the use of influent models requires a deep understanding of the underlying system and its intricacies.

Phung et al. [17] found that the fault diagnosis performance using estimated values by the auto-associative neural network (ANN) with a recursive minimization strategy would notably decrease when the percentage of missing values surpasses around 30%. However, if the missing values are not predominant across most variables for each sample simultaneously, the acceptable limit for the percentage of missingness could be slightly higher.

Zhang et al. [11] asserted that the self-organizing map (SOM) method requires a large amount of data to be effective. In a similar vein, Rustum & Adeloye in [3] emphasized that the proposed KSOM need a large amount of data to train the KSOM. Additionally, the KSOM exhibits sensitivity to initial conditions and poses challenges in determining the appropriate number of nodes [3]. SOM method may not work well with categorical or binary data, as it is designed for continuous and numeric variables. Also, the accuracy of the imputed values depends on the quality of the training data and the relationships between variables [25]. It is also proved that the KSOM is not suitable for predicting extreme values that are outside the range of the training data [3].

According to Huo et al. [10], they observed that the TES and TESWN methods depend on time series models, which might not produce satisfactory outcomes when dealing with missing data unrelated to time. Furthermore, the TES method neglects the uncertainty associated with the missing value, resulting in an underestimation of the population variance for both influent data and simulated effluent concentrations. Huo et al. [9] It was stated that the TES method functions based on the assumption that the missing values are missing at random (MAR). Consequently, the accuracy of the imputed values may be influenced by the assumptions made by the method. Additionally, the TES method could erroneously introduce abrupt temporal changes in variables within the data record. In fact, the performance of TES and TESWN methods may be contingent on specific characteristics of the data being imputed, such as the degree of autocorrelation and the presence of outliers [10].

According to the findings of Villez et al. [12], in the missing data imputation technique using PCA, the undetected failure ratio appears to be significantly impacted by the presence of gaps in the data. This ratio nearly doubles when no estimates are employed. This suggests this method may not be able to accurately estimate missing data in all cases, leading to potential limitations in the performance of the model.

In their paper, Oliveira-Esquerre et al. [24] applied linear interpolation to estimate missing values, but they limited it to no more than two consecutive missing values. This indicates that linear interpolation might not be effective for estimating missing values when there are more than two consecutive missing values. The original database covered a period of 1427 consecutive days, roughly a four-year daily record. However, the significantly high occurrence of missing values for several variables, particularly for TSS, NAM, and NN variables, poses a substantial issue in the dataset. Missing values are more prevalent than available data for these variables.

IV. Discussion

Findings from this paper reveal that there are only a limited number of articles (16 in total) discussing imputation methods used in WWTPs. This suggests that the literature on this filed is relatively scarce. However, despite the limited number of studies, the findings indicate a diverse range of approaches being explored.

These studies focused on various imputation methods, including the MSF-ARI approach (R2AU-Net) [20], univariate imputation methods (such as the SSD method and SVR model) [21], the SVM algorithm [16], coupling the expectation maximization (EM) algorithm with the Markov Chain Monte Carlo (MCMC) method (EM-MCMC) [22], Case-Based Reasoning (CBR) approach [18], Cubic Hermite interpolation method [14], Non-Linear decreasing inertia weight particle swarm algorithm (NLDIW-PSO) based optimal Support Vector Regression (SVR) [23], Daily average, auto-associative neural network (ANN) with a recursive minimization strategy [19], Multiple imputation (MI) [17], Linear interpolation [24], Two-directional exponential smoothing (TES) [9], PCA projection method [12], Kohonen self-organizing map (KSOM) [3], TES and TESWN [10].

The effectiveness of these imputation methods in handling missing data challenges was evaluated based on several criteria, including imputation accuracy, computational efficiency, robustness to different types and patterns of missing data [9][10]. The results varied across the studies, with some methods demonstrating high accuracy in imputing missing values [16][18][23], while others showed limitations in certain aspects [3] [11] [12] [14] [22].

Several limitations were identified across the reviewed papers. One common limitation was the lack of generalizability of proposed approaches to different WWTPs with varying configurations and operating conditions. In some cases, the proposed methods required a substantial amount of training data [3] [11], which may not be available in all WWTPs. Another limitation was the failure to consider sensor drift, which could impact the accuracy of imputed data. Furthermore, many studies did not compare their proposed methods with other state-of-the-art imputation techniques or evaluate their performance on diverse datasets. The generalizability of findings was often limited by the use of data from a single WWTP or a specific location [3] [9] [24], raising concerns about the applicability of the proposed methodologies to other systems.

Future research should address these limitations by conducting broader investigations, comparing with existing methods, and exploring the impact of various factors on data imputation in WWTPs to bridge the existing knowledge gaps and ensure the reliable management and analysis of data in this domain.

V. Conclusion

In this study, the SLR examine the existing missing data imputation methods used in WWTPs. This SLR is also concerned with aiding researchers working in this field in the decision-making processes and enhancing the performance of WWTPs. This study concentrated on the scientific database Scopus.

The findings from the selected studies reveal a limited number of articles discussing this specific topic with only 16 articles meeting the inclusion criteria. Despite the scarcity of literature, the findings demonstrate a diverse range of approaches being explored in this field.

The studies indicate that these imputation methods have shown promising results in handling missing data in various aspects of WWTPs, including influent data and water quality data. They have been employed to impute missing values for different variables, such as flow, temperature, BOD5, suspended solids, ammonia nitrogen, pH values, and more. The effectiveness of these methods has been evaluated using different evaluation metrics, such as mean squared error (MSE), MAE, and coefficient of determination (R^2).

However, it is important to acknowledge some limitations identified in the reviewed studies. These include the lack of generalizability of proposed approaches to different WWTPs with varying configurations and operating conditions, the requirement for a substantial amount of training data which may not be universally available, the failure to consider sensor drift in imputation methods and the need for comparing the proposed methods with other state-of-the-art techniques are also areas that require attention.

Based on these findings, several recommendations can be made. Further investigation is warranted, specifically taking into account the missing mechanisms and rates associated with data gaps in this particular field. Furthermore, researchers should aim to validate and generalize the proposed imputation methods by conducting experiments in multiple WWTPs with diverse characteristics. This will enhance the understanding of their performance and applicability in different settings. Moreover, additional evaluation metrics such as RMSE and MAE should be employed to comprehensively assess the effectiveness of the imputation methods. Comparative studies, benchmarking the proposed methods against other state-of-the-art techniques, would also provide valuable insights into their relative strengths and weaknesses.

References

[1] C. Rosen, “Monitoring wastewater treatment system.,” thesis, Dept. of Industrial Electrical Engineering and Automation, Lund Institute of Technology, Lund Univ., Lund, Sweden., 1998.

[2] J. Huo, C. Cox, W. Seaver, B. Robinson, and Y. Jiang, “Innovative missing data replacement methods using time series models,” in World Environmental and Water Resources Congress 2008: Ahupua’a - Proceedings of the World Environmental and Water Resources Congress 2008, 2008. doi: 10.1061/40976(316)670.

[3] R. Rustum and A. J. Adeloye, "Replacing outliers and missing values from activated sludge data using kohonen self-organizing map," Journal of Environmental Engineering, vol. 133, no. 9, pp. 909–916, 2007, doi: 10.1061/(ASCE)0733-9372(2007)133:9(909).

[4] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, Dec. 1976, doi: 10.1093/BIOMET/63.3.581.

[5] B. L. Ford, "An overview of hot-deck procedures," Incomplete Data Sample Surveys, vol. 2, 1983.

[6] N. Z. Zainal Abidin, A. R. Ismail, and N. A. Emran, "Performance Analysis of Machine Learning Algorithms for Missing Value Imputation," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, 2018.

[7] L. Bargelloni et al., "Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream," Aquac. Reports, vol. 20, pp. 100–661, 2021.

[8] L. Yang et al., "Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective," Inf Softw Technol, vol. 130, p. 106397, Feb. 2021, doi: 10.1016/J.INFSOF.2020.106397.

[9] J. Huo, W. L. Seaver, R. B. Robinson, and C. D. Cox, "Application of Time Series Models to Analyze and Forecast the Influent Components of Wastewater Treatment Plants (WWTPs)," World Water Congress 2005: Impacts of Global Climate Change - Proceedings of the 2005 World Water and Environmental Resources Congress, pp. 1–11, 2005, doi: 10.1061/40792(173)97.

[10] J. Huo, C. D. Cox, W. L. Seaver, R. B. Robinson, and Y. Jiang, "Application of two-directional time series models to replace missing data," Journal of Environmental Engineering, vol. 136, no. 4, pp. 435–443, 2010, doi: 10.1061/(ASCE)EE.1943-7870.0000171.

[11] L. Zhang, M. Scholz, A. Mustafa, and R. Harrington, "Application of the self-organizing map as a prediction tool for an integrated constructed wetland agroecosystem treating agricultural runoff," Bioresour Technol, vol. 100, no. 2, pp. 559–565, Jan. 2009, doi: 10.1016/J.BIORTECH.2008.06.042.

[12] K. Villez, C. Rosen, V. H. Stijn, C. K. Yoo, and P. A. Vanrolleghem, "On-line dynamic monitoring of the SHARON process for sustainable nitrogen removal from wastewater," Computer Aided Chemical Engineering, vol. 20, no. C, pp. 1297–1302, Jan. 2005, doi: 10.1016/S1570-7946(05)80058-6.

[13] P. R. C. P. A. T. and J. F. M. Nelson, "Missing methods in PCA and PLS: Score calculations with incomplete observations," Chem. Intell. Lab. Syst., vol. 35, p. 45, 1996.

[14] S. Borzooei, G. H. B. Miranda, R. Teegavarapu, G. Scibilia, L. Meucci, and M. C. Zanetti, "Assessment of weather-based influent scenarios for a WWTP: Application of a pattern recognition technique," J Environ Manage, vol. 242, pp. 450–456, Jul. 2019, doi: 10.1016/J.JENVMAN.2019.04.083.

[15] C. De Mulder, T. Flameling, S. Weijers, Y. Amerlinck, and I. Nopens, "An open software package for data reconciliation and gap filling in preparation of Water and Resource Recovery Facility Modeling," Environmental Modelling & Software, vol. 107, pp. 186–198, Sep. 2018, doi: 10.1016/J.ENVSOFT.2018.05.015.

[16] F. Azizoğlu and E. Ünsal, "Kayıp IoT Verilerinin Makine Öğrenmesi Teknikleri ile Tahmini," El-Cezeri, vol. 9, no. 4, pp. 1388–1397, Dec. 2022, doi: 10.31202/ECJSE.1135485.

[17] D. Phung et al., "Can wastewater-based epidemiology be used to evaluate the health impact of temperature? – An exploratory study in an Australian population," Environ Res, vol. 156, pp. 113–119, Jul. 2017, doi: 10.1016/J.ENVRES.2017.03.023.

[18] J. Pascual-Pañach, M. Sànchez-Marrè, and M. À. Cugueró-Escofet, "Optimizing Online Time-Series Data Imputation Through Case-Based Reasoning," Frontiers in Artificial Intelligence and Applications, vol. 356, pp. 87–90, Oct. 2022, doi: 10.3233/FAIA220320.

[19] H. Xiao, D. Huang, Y. Pan, Y. Liu, and K. Song, "Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model," Chemometrics and Intelligent Laboratory Systems, vol. 161, pp. 96–107, Feb. 2017, doi: 10.1016/J.CHEMOLAB.2016.12.009.

[20] A. H. Ba-Alawi, K. J. Nam, S. K. Heo, T. Y. Woo, H. Aamer, and C. K. Yoo, "Explainable multisensor fusion-based automatic reconciliation and imputation of faulty and missing data in membrane bioreactor plants for fouling alleviation and energy saving," Chemical Engineering Journal, vol. 452, p. 139220, Jan. 2023, doi: 10.1016/J.CEJ.2022.139220.

[21] H. Han, M. Sun, H. Han, X. Wu, and J. Qiao, "Univariate imputation method for recovering missing data in wastewater treatment process," Chin J Chem Eng, vol. 53, pp. 201–210, Jan. 2023, doi: 10.1016/J.CJCHE.2022.01.033.

[22] H. Safford et al., "Wastewater-Based Epidemiology for COVID-19: Handling qPCR Nondetects and Comparing Spatially Granular Wastewater and Clinical Data Trends," ACS ES and T Water, vol. 2, no. 11, pp. 2114–2124, Nov. 2022, doi: 10.1021/ACSESTWATER.2C00053/SUPPL_FILE/EW2C00053_SI_005.XLSX.

[23] J. Yan, X. Chen, and Y. Yu, "A Data Cleaning Framework for Water Quality Based on NLDIW-PSO Based Optimal SVR," Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018, pp. 336–341, Jan. 2019, doi: 10.1109/WI.2018.00-71.

[24] K. P. Oliveira-Esquerre, D. E. Seborg, R. E. Bruns, and M. Mori, "Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill: Part I. Linear approaches," Chemical Engineering Journal, vol. 104, no. 1–3, pp. 73–81, Nov. 2004, doi: 10.1016/J.CEJ.2004.05.011.

[25] L. Zhang, M. Scholz, A. Mustafa, and R. Harrington, "Application of the self-organizing map as a prediction tool for an integrated constructed wetland agroecosystem treating agricultural runoff," Bioresour Technol, vol. 100, no. 2, pp. 559–565, 2009, doi: 10.1016/j.biortech.2008.06.042.