Optimizing wastewater treatment plants with advanced feature selection and sensor technologies

MÍRIAM TIMIRAOS[*], *Department of Industrial Engineering, University of A Coruña, CTC, Ferrol, 15071 A Coruña, Spain; Department of Water Technologies, Fundación Instituto Tecnológico de Galicia, National Technological Center, 15003, A Coruña, Spain.*

JESÚS F. ÁGUILA[**], *Department of Water Technologies, Fundación Instituto Tecnológico de Galicia, National Technological Center, 15003, A Coruña, Spain.*

ELENA ARCE[†], *Department of Industrial Engineering, University of A Coruña, CTC, Ferrol, 15071 A Coruña, Spain.*

MOISÉS ALBERTO GARCÍA NÚÑEZ[††], *Department of Enterprise, Faculty of Labour Sciences, University of A Coruña, Calle San Ramón s/n, 15403, A Coruña, Spain.*

FRANCISCO ZAYAS-GATO[§], *Department of Industrial Engineering, University of A Coruña, CTC, Ferrol, 15071 A Coruña, Spain; University of A Coruña, CITIC, Campus de Elviña, 15071 A Coruña, Spain.*

HÉCTOR QUINTIÁN[§§], *Department of Industrial Engineering, University of A Coruña, CTC, Ferrol, 15071 A Coruña, Spain; University of A Coruña, CITIC, Campus de Elviña, 15071 A Coruña, Spain.*

Abstract

This research establishes a foundational framework for the development of virtual sensors and provides significant preliminary results. Our study specifically focuses on identifying the key factors essential for accurately predicting total nitrogen in the effluent of wastewater treatment plants. This contribution enhances the predictive capabilities and operational

[*] E-mail: miriam.timiraos.diaz@udc.es
[**] E-mail: jfernandez@itg.es
[†] E-mail: elena.arce@udc.es
[††] E-mail: moises.alberto.garcia@udc.es
[§] E-mail: f.zayas.gato@udc.es
[§§] E-mail: hector.quintian@udc.es

Vol. 00, No. 0, © The Author(s) 2024. Published by Oxford University Press.

https://doi.org/10.1093/jigpal/jzae108

efficiency of these plants, demonstrating the practical benefits of integrating advanced feature selection methods and innovative sensor technologies. These findings provide crucial insights and pave the way for future advancements in the field. In this study, four different feature selection methods are employed to comprehensively explore the variables influencing total nitrogen predictions. The effectiveness of these methods is then evaluated by applying three regression techniques. The findings indicate acceptable levels of accuracy in all applied cases, with one method demonstrating particularly promising results, applicable to several wastewater treatment plants. This validation of the selected variables not only underlines their effectiveness, but also lays the foundation for future virtual sensor applications. The integration of such sensors promises to improve the accuracy and reliability of predictions, marking a significant advance in wastewater treatment plant instrumentation.

Keywords: Feature selection, wastewater treatment plant, regression techniques, prediction, total nitrogen

1 Introduction

The contemporary reality of water shortage due to factors like climate change is indisputable [43]. Concurrently, the global population continues to grow [9], and with an elevated standard of living, there is an observable surge in water consumption [7]. Unsurprisingly, heightened water consumption results in increased wastewater production. An increment of wastewater production means that treatment plants are often unable to cope with the growth energy and resources required to process these large quantities of water [42].

In light of these circumstances, optimizing the operation of wastewater treatment plants (WWTPs) becomes imperative [45]. Numerous efforts have been made to enhance the efficiency of these facilities through various approaches [16]. Depending on the type of sewer network, these plants may receive domestic and industrial wastewater in the case of separative sewer networks, or they may also include rainwater when a single piping system transports all types of water generated in the population center (combined sewer networks) [27]. The level of pollution in the wastewater to be treated by them in the latter scenario is usually lower during rainfall events, although the volume of water tends to increase significantly, sometimes exceeding the treatment capacity of the facilities [16, 27].

The particular characteristics of each WWTP and its location, as well as the temporal variations of the quantity and quality of the wastewater to be treated, require a high level of monitoring in the different treatment processes to optimize their operation and to meet the water purification requirements established by regulations [6, 25]. Monitoring in them is essential for multiple reasons: 1) real-time control of different treatment processes through tracking key parameters, 2) optimization of energy consumption by adjusting processes according to treatment demands, 3) early detection of anomalies and implementation of predictive maintenance strategies, 4) optimization of sludge generation and management, 5) adaptation of the plant operation based on changes in the pollutant load of the incoming raw wastewater due to climatic factors or temporal patterns in water consumption and 6) control the quality of the treated water discharged to water bodies. Despite the associated advantages in the efficient management of WWTPs, the initial economic investment to achieve a high level of monitoring in all treatment processes is usually significant. Therefore, minimizing the number of sensors to be installed, identifying the most relevant variables to be measured or implementing virtual sensors plays a crucial role in optimizing the management and cost savings derived from the operation of WWTPs [19, 24, 26]. For example, monitoring parameters such as total nitrogen in these plants to quantify the amount of organic matter, proteins and amino acids present in the water plays an important role in understanding the overall degree of contamination and serves as an indicator to optimize various treatment processes.

Furthermore, the measurement of this variable involves considerations beyond a simple sensor, requiring focus, experience and specific measurement methods to obtain reliable results [3, 50].

The integration of new sensor technologies into treatment plants offers significant advancements in operational efficiency and monitoring capabilities. These technologies, such as advanced optical sensors, biosensors and IoT-enabled devices, enable real-time tracking and data collection of various parameters crucial for effective wastewater treatment. For instance, optical sensors can provide continuous measurements of pollutants like nitrogen and phosphorus, while biosensors can detect specific biological markers indicative of water quality issues [3, 50].

Implementing these technologies in day-to-day operations facilitates early detection of anomalies, predictive maintenance and adaptive process control, which collectively enhance the overall treatment efficiency. Practical application of these sensors requires addressing challenges such as integration with existing systems, ensuring data accuracy and reliability and managing the economic investment associated with their deployment. By incorporating new sensor technologies, WWTPs can achieve better compliance with environmental regulations, optimize energy usage and improve the quality of treated effluent, ultimately contributing to sustainable water management practices [3, 50].

The exploration of strategies to enhance the energy efficiency of treatment plants is a critical area of research, as demonstrated by a study presented in [8]. This investigation delves into various options aimed at optimizing the energy consumption of Italy's largest facility. As energy consumption is a significant operational cost for such facilities, efficiency improvements can lead to substantial economic benefits.

In a complementary modeling approach, discussed in [31, 36], the determination of the optimal solid retention time represents a key aspect of reducing operating expenses effectively. Solid retention time optimization contributes to the efficient removal of pollutants from wastewater, thereby streamlining the treatment process and enhancing cost-effectiveness.

Furthermore, the refinement of ozonation processes by eliminating standard substances is addressed in [12]. This approach is pivotal in improving the overall treatment efficiency of WWTPs, as it targets the optimization of a specific treatment step, ultimately leading to more effective pollutant removal and operational cost reduction.

Several studies support the notion that optimizing WWTPs yields tangible benefits. For instance, research highlighted in [42, 52] emphasizes the efficacy of optimization strategies in reducing operating costs and enhancing the overall efficiency of wastewater treatment processes. These findings underscore the importance of continuous efforts to refine and optimize the operation of them.

Given that many wastewater treatment facilities are publicly owned, the imperative of cost optimization becomes even more pronounced. Publicly funded facilities must operate efficiently to ensure responsible resource allocation and effective waste management. Moreover, the significance of treated water as a valuable resource is underscored, particularly in regions confronting frequent droughts [45]. This highlights the dual benefit of wastewater treatment optimization, not only in terms of cost reduction but also in the sustainable production of a valuable water resource, addressing challenges posed by water scarcity in drought-prone areas.

In [10] a novel technique is proposed to monitor the presence of foam in WWTP tanks in real time using texture segmentation models trained with centralized and federated approaches. The proposed methodology is integrated into an image processing chain that consists of capturing images using a professional camera, ensuring the absence of anomalies in the captured images and implementing a real-time communication method for event notifications to plant operators.

Although the aforementioned studies [8, 12, 31, 36, 42, 52] have made interesting contributions to the field by exploring strategies such as improving energy efficiency, determining the optimal solids retention time and refining the ozonation process, further research is still needed to overcome certain limitations. For example, the focus of the study on the largest plant in Italy may limit the generalizability of the results to other geographical and operational contexts. In addition, the optimization strategies proposed in [31, 36] and [12] may not fully take into account the multiple challenges faced by these plants, such as the prediction of parameters of great interest like total nitrogen.

In light of these considerations, this work introduces a novel method for identifying representative variables in WWTPs, which is a critical step towards optimizing the complex and dynamic nature of wastewater treatment in future studies. Unlike previously mentioned studies, the proposed approach emphasizes the importance of continuous monitoring across an extensive set of parameters, allowing robust feature selection to significantly reduce the number of sensors required. This reduction not only minimizes the economic investment associated with installing a large number of sensors, but also ensures that only the most relevant variables, crucial for regulatory compliance, are retained.

In addition, our work employs regression techniques to identify relevant variables that influence the output variable, total nitrogen, setting the stage for future predictive modeling. This approach lays the groundwork for the potential creation of virtual sensors, which could provide a cost-effective and efficient means of monitoring and controlling the treatment process in the future. By identifying key variables and establishing validation mechanisms, we aim to add a layer of reliability to the monitoring system.

This article presents a method for identifying representative variables in a WWTP, as well as the construction of a prediction model for the selected ones. Numerous parameters are monitored during plant operation, to apply feature selection to reduce the required number of sensors significantly. The goal is to retain only the sensors strictly necessary for regulatory compliance. Regression techniques are used to create adjusted models for the prediction of the output variable, which allows the creation of virtual sensors, as well as a mechanism to check real measurements coming from the sensor.

The document is structured as follows: after this introduction, the case study is presented. Subsequently, the methods employed are explained, followed by describing experiments and results. Finally, conclusions are drawn, and future works are proposed.

2 Case of study

A WWTP is a set of facilities, typically located on the outskirts of or outside population centers, whose main function is to reduce the pollution of wastewater to acceptable limits for discharge into the aquatic environment. Depending on their size and the pollutants to be treated, they are comprised of different treatment processes organized into distinct operational lines. In general terms, two main operation lines are usually distinguished: the water line, focused on wastewater purification, and the sludge line, centering on the management of solids (sludges) generated in the treatment processes [39]. This research has been accomplished over a medium-sized WWTP located in a Mediterranean climate site that serves an area with a population of around 15000 people. Figure 1 shows an overview of the processes of the WWTP used in this study, encompassing the main wastewater treatments from the inflow of raw wastewater into the plant to the discharge of the treated effluent to the aquatic environment. The water line consists mainly of pretreatment, secondary treatment (including anoxic and aerobic phases) and tertiary treatment stages.

Raw wastewater enters the water line to undergo preliminary treatment, aiming to remove coarse and fine solids in the screening stage, as well as greases and oils in the grit and grease removal

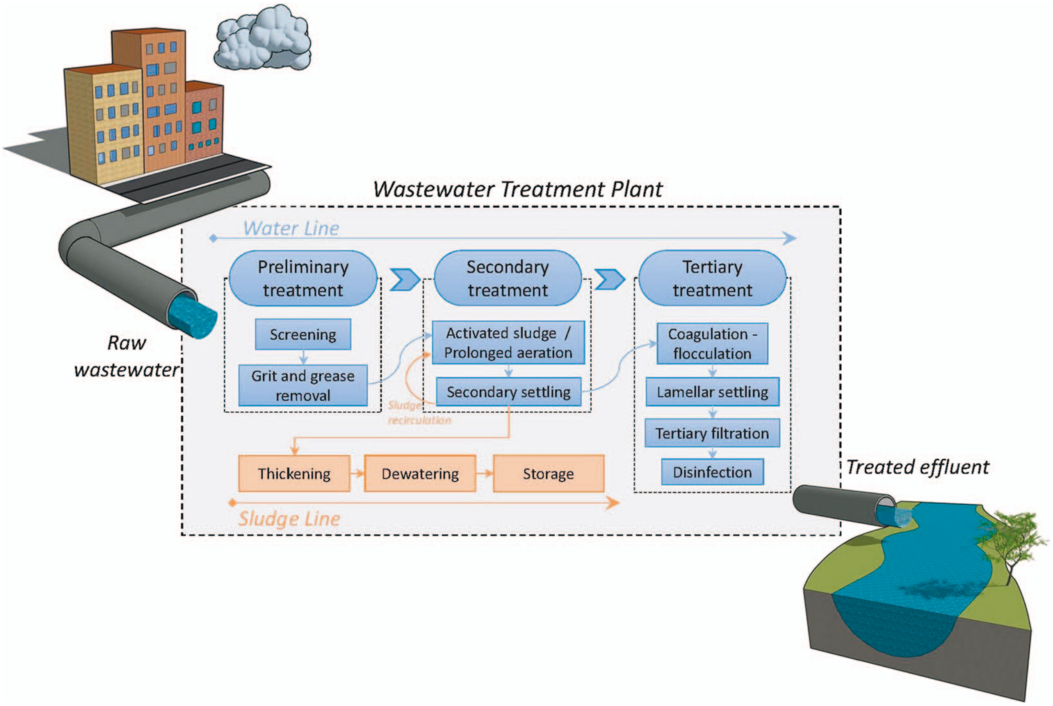


FIGURE 1. Diagram of the WWTP under investigation, illustrating the main treatment processes in the water operation line (preliminary treatment, secondary treatment and tertiary treatment) and the sludge operation line, from the raw wastewater inlet to the discharge of the treated effluent into the aquatic environment.

system [33]. The resulting wastewater undergoes secondary treatment for the removal of dissolved and suspended organic matter remaining after preliminary treatment. The treatment plant under study performs this process by applying activated sludge technology and prolonged aeration [34]. In this stage, the treatment process is carried out combining an anoxic phase without oxygen supply to the system to favor nitrate removal and an aerobic phase in which oxygen is supplied to aerobic microorganisms responsible for breaking down organic matter [5]. The wastewater treatment continues with secondary settling to separate the solid waste or sludge formed during the previous biological processes [35]. The water line concludes with tertiary treatment to achieve a higher degree of removal of specific contaminants, such as phosphorus, still present in the water from secondary settling [51]. In the analysed facility, physicochemical treatments of coagulation–flocculation, lamellar settling and filtration are carried out [28]. Tertiary treatment concludes with a disinfection process to eliminate pathogenic microorganisms present in the treated water through exposure to ultraviolet rays and the addition of sodium hypochlorite. Once the raw water reaches the plant and has been treated by passing through all the processes of the water line, it is discharged into a water body. Part of the sludge generated during wastewater treatment is collected from the bottom of the secondary settling tanks of the WWTP to be recirculated to the biological reactor and become part of the biological concentrate. At the same time, the excess produced will move to the sludge line for treatment. This sludge often contains a significant amount of water, so it undergoes a gravity-

TABLE 1. Variables in the dataset.

Input	Variable Name
pH at the Entrance	PH_E
pH on Exit	PH_S
Conductivity at the Entrance	Conductivity_E
Conductivity at the Exit	Conductivity_S
V60 at the Entrance	V60_E
Solids in Suspension at the Entrance	SS_E
Solids in Suspension on Exit	SS_S
Biological Oxygen Demand on Exit	BOD_S
Chemical Oxygen Demand at the Entrance	COD_E
Chemical Oxygen Demand on Exit	COD_R
Total Nitrogen at the Entrance	Nitrogen_T_E
Total Phosphorus at the Entrance	Phosphorus_T_E
Total Phosphorus on Exit	Phosphorus_T_S
Ammonia at the Entrance	NH3_E
Total Kjeldahl Nitrogen at the Entrance	NTK_E
Nitrate at the Entrance	NO3_E
Nitrogen dioxide at the Entrance	NO2_E
Ammonia on Exit	NH3_S
Total Kjeldahl Nitrogen on Exit	NTK_S
Nitrate on Exit	NO3_S
Nitrogen dioxide on Exit	N02_S
Thickener input	Input_Esp
Output	**Variable Name**
Total Nitrogen on Exit	Nitrogen_T_S

thickening process where the solids concentration is increased, facilitating its handling and further processing [2]. After thickening, the sludge undergoes a centrifugal dewatering to further reduce the water content. Finally, the sludge is temporarily stored at the WWTP until its final transfer for various uses, such as agricultural fertilizers [11].

In this research work, a dataset consisting of 23 monitored variables in the plant has been analysed. The samples composing it have been collected for nine months, with a recording frequency of one value per day. In Table 1, the variables available in the utilized dataset are displayed.

The variables listed in Table 1 are monitored at both the inlet and outlet of the WWTP. However, specific details regarding the exact location of each variable within the plant are not available at this stage of the research. Future studies may provide more detailed information on the specific monitoring locations for each variable.

3 Applied methods

In this research, feature selection methods are applied in conjunction with regression algorithms to identify the relationship between variables measured at the inlet and outlet of a WWTP. The methodology is divided into two stages. The first stage is aimed at assessing the significance of the input variables with respect to the output variable. Subsequently, in the second stage, the identified correlations are verified, and a regression model is constructed.

3.1 Feature selection

This document focuses on selecting relevant characteristics by evaluating their relevance and redundancy. Features are classified as: 1) highly relevant, 2) not very relevant but not redundant, 3) irrelevant and 4) redundant. Highly relevant features are crucial and cannot be removed without affecting the original distribution. Not very relevant but not redundant are variables that do not have a high relevance, but cannot be forgotten as they could be necessary. Instead, depending on specific conditions, irrelevant features may not be necessary, and redundant features can be replaced without affecting the distribution. The objective is to increase relevance and minimize redundancy by searching for a subset of only relevant features [18, 23, 41].

There are different methods for feature selection, including filter, wrapper, embedding and hybrid methods [18, 49]. These methods assume that features are independent or almost independent. However, other methods exist for datasets that contain structured features with dependencies and flow features. For this article, the focus will be on discussing the methods included in the common classification [22, 49].

Filter methods rank features based on statistical measures and are generally faster but may overlook interactions between features [18, 49]. Wrapper methods use a predictive model to evaluate feature subsets, potentially offering higher accuracy but at the cost of greater computational complexity [18, 49]. Embedding methods incorporate feature selection within the training process of the predictive model, balancing performance and computational efficiency. Hybrid methods combine aspects of both filter and wrapper approaches, aiming to leverage the strengths of each [18, 49].

For instance, the application of Principal Component Analysis (PCA) can reduce dimensionality and eliminate redundancies, enhancing computational efficiency, though it might miss some nonlinear relationships. In contrast, machine learning-based methods like Random Forest can capture complex interactions between variables but require more computational resources. Understanding these trade-offs helps in selecting the most appropriate method for specific scenarios, improving the accuracy and reliability of total nitrogen predictions in WWTPs [22, 49].

3.1.1 Correlation Matrix The correlation matrix, denoted by R, is a fundamental component in statistical analysis and plays a pivotal role in understanding the relationships between variables. It is a square and symmetric matrix that provides a comprehensive overview of the pairwise correlations among variables within a dataset [17, 29].

At its core, the correlation matrix serves as a powerful tool for examining the strength and direction of linear relationships between variables. Each element in the matrix represents the correlation coefficient between two variables, indicating the extent to which they are associated. The values range from -1 to 1, where 1 signifies a perfect positive correlation, -1 denotes a perfect negative correlation and 0 indicates no linear correlation [17].

The main diagonal of the correlation matrix is always populated with 1s, as it represents the correlation of a variable with itself, which is perfect. The off-diagonal elements contain the correlation coefficients between pairs of distinct variables. The symmetry of the matrix arises from the fact that the correlation between variable A and B is the same as the correlation between B and A [29].

Including a visual representation of the correlation matrix can further elucidate its structure 1.

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} & .... & r_{1p} \\ r_{21} & 1 & r_{23} & r_{24} & .... & r_{2p} \\ r_{31} & r_{32} & 1 & r_{34} & .... & r_{3p} \\ r_{p1} & r_{p2} & r_{p3} & 1 & .... & 1 \end{pmatrix} \tag{1}$$

3.1.2 LASSO The primary objective is to pinpoint the most influential variables and their associated regression coefficients that contribute to the development of a model characterized by minimal prediction error. This objective is accomplished through the application of a constraint on the model parameters, a mechanism designed to encourage regression coefficients to converge towards zero [13, 32, 37].

In the pursuit of identifying critical variables, the modeling process involves scrutinizing the impact of each variable on the predictive performance of the model. The emphasis is on discerning the variables that significantly contribute to minimizing prediction errors, thus enhancing the overall accuracy and reliability of the predictive model [32].

One widely employed method to achieve this goal is the imposition of a regularization technique, such as Lasso or Ridge regression. These methods introduce a penalty term to the traditional regression model, effectively constraining the magnitude of the regression coefficients. By penalizing large coefficients, the regularization process encourages the model to favor simpler solutions, often resulting in coefficients approaching or equaling zero [37].

The mathematical representation of this concept involves modifying the standard regression equation with a regularization term. For instance, in Lasso regression, the modified objective function is expressed as 2 [13].

$$Minimize : \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \Rightarrow Minimize : RSS + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{2}$$

3.1.3 Mutual Info Regression Mutual information regression serves as a quantifiable metric for assessing the relationship between two random variables, offering insights into the degree of dependence between them. This measure inherently assumes non-negativity, meaning it only yields values greater than or equal to zero. In the context of mutual information regression, the value of zero signifies independence between the two variables, with higher values indicative of a more pronounced and stronger dependency [40].

To elaborate further, mutual information (MI) quantifies the amount of information shared between two variables. If the two variables under consideration are independent, meaning changes in one variable do not provide any information about the other, the mutual information will be zero. This scenario reflects the absence of shared information, signifying independence in the observed data. Conversely, as the MI value increases, it denotes an elevated level of dependency between the variables. Higher values suggest that changes in one variable are associated with corresponding changes in the other, indicating a stronger relationship. This dependency could manifest as a linear correlation, non-linear association or some other form of statistical dependence, depending on the nature of the variables and their relationship [20].

3.1.4 Random Forest The Random Forest algorithm represents a sophisticated ensemble approach in machine learning, adept at generating decision trees through a technique known as bootstrap aggregation. The culmination of this process is a collective prediction derived from the amalgamation of predictions made by the individual trees. A distinctive characteristic of the Random Forest methodology is its deliberate consideration of all possible features during the tree-building process, coupled with a mechanism to mitigate the issue of high correlation among the constituent trees [4, 18].

The key innovation in Random Forest lies in the concept of bootstrap aggregation, commonly referred to as bagging. This involves creating multiple decision trees by drawing random samples, with replacements, from the original dataset. Each tree is constructed independently, introducing diversity into the ensemble. Subsequently, the predictions from all trees are combined to form a more robust and accurate overall prediction [21].

Noteworthy is the algorithm's meticulous consideration of all available features when growing individual trees. During the construction of each decision tree, Random Forest assesses potential splits at each node using a random subset of features. This randomness ensures that each tree in the ensemble explores different aspects of the dataset, contributing to a diverse set of predictions [21].

Furthermore, to prevent the trees within the Random Forest ensemble from becoming highly correlated, an additional layer of randomness is introduced. At each node of the decision tree, a random subset of features is considered for split candidates, contributing to decorrelating the trees and enhancing the overall predictive power of the ensemble [4].

3.2 Regression techniques

Regression analysis refers to a collection of statistical techniques used to measure associations between a dependent variable and one or more independent variables. Its main goal is to evaluate the strength of the connections between these variables and create models that can predict future relationships between them [1, 14].

In selecting regression methods for our study, it is carefully considered various factors, including the nature of the data and the suitability of each method for our research objectives. The chosen methods, including K-Nearest Neighbors (KNN), Linear Regression (LR) and Decision Trees (DT), were selected for their unique strengths and applicability to our specific context. Each method offers distinct advantages, such as capturing nonlinear relationships (KNN), simplicity and interpretability (LR) and handling nonlinear relationships and interactions (DT). Together, these methods provide a comprehensive framework for analysing the relationships between variables.

3.2.1 Linear Regression (LR) Linear Regression (LR) stands as a foundational statistical technique widely employed for predictive modeling, specifically designed to estimate the value of a variable based on the information provided by one or more independent variables. This method is particularly adept at establishing a linear relationship between the dependent variable and the chosen independent variables, allowing for the formulation of a predictive model in the form of a linear equation [38].

The central objective of LR is to determine the coefficients of the linear equation, thereby establishing the relationship between the independent and dependent variables. The linear equation takes the form in 3 where Y is the independent variable, X_1, X_2, X_n are the dependent variables, β_0 is the intercept, $\beta_1, \beta_2, \beta_n$ are the coefficients representing the impact of each independent variable and ϵ is the error term accounting for unexplained variability [46].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{3}$$

The versatility of LR extends across diverse fields of study, including but not limited to, economics, finance, biology and, as mentioned in the provided reference [38, 46], even in the context of WWTP. In the realm of wastewater treatment, LR models can be employed to predict key variables such as pollutant concentrations or treatment efficiency based on various input parameters. This predictive capability aids in optimizing operational processes and ensuring the effective management of wastewater treatment facilities.

3.2.2 K-Nearest Neighbors (KNN) KNNs emerge as a non-parametric regression method, offering a flexible approach to predicting continuous outcomes based on the relationships among independent variables. Unlike parametric models such as linear regression, KNN does not assume a specific functional form for the underlying relationship, making it well-suited for scenarios with complex or non-linear patterns. The core principle of KNN involves estimating the correlation between independent variables and the continuous target by computing the average of observations within the same neighborhood. The prediction formula for KNN can be represented as 4, where $\hat{y}$ represents the predicted value, y_i represents the value of the target variable for the $i - th$ nearest neighbor and k represents the number of neighbors [53].

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{4}$$

In the KNN framework, each data point in the dataset becomes a potential predictor, and predictions for a new data point are determined by examining its proximity to the neighboring points in the feature space. The ‘k’ in KNN signifies the number of nearest neighbors considered when making predictions. The average or weighted average of the target variable in these KNNs serves as the predicted value for the new data point [30].

The versatility of KNN renders it applicable across diverse domains, ranging from healthcare and finance to environmental science. It is particularly advantageous when dealing with datasets exhibiting intricate patterns or irregularities that may not be well-captured by traditional parametric models [30, 53].

3.2.3 Decision Trees (DT) DTs stand as a robust and versatile algorithm within the realm of machine learning, renowned for its capacity to effectively partition data into groups based on their categories or output values. This algorithm embodies a recursive process of data division, enabling it to hierarchically organize and classify information, ultimately facilitating accurate predictions [48].

The fundamental learning mechanism of DT involves a two-step process: data partitioning and prediction. Initially, the algorithm strategically divides the training data into subsets at each node of the tree. The objective of this partitioning is to minimize the sum of squared residuals, emphasizing the reduction of variability within each subgroup. This process continues iteratively, with each subsequent split aiming to enhance the homogeneity of data within the resulting nodes [47].

Upon completion of the partitioning phase, DT harness the acquired knowledge to make predictions. For a given input, the algorithm traverses the tree structure, following the decision rules established during training. The prediction is then made based on the majority class or the average response of the instances falling into the corresponding leaf node. The prediction formula for a Decision Tree involves traversing the tree structure to reach a leaf node corresponding to the predicted value. While the exact mathematical formulation varies depending on the specific implementation, the decision-making process involves evaluating feature splits at each node to determine the path through the tree [47, 48].

4 Experiments and results

This paper evaluates the performance of combining four feature selection methods (Matrix Correlation, Random Forest, Mutual Info Regression and LASSO) with three regression techniques (Linear Regression, KNN and DT) to evaluate the correlation between a set of input variables with an output variable, the total nitrogen at the outlet of a WWTP. For this purpose, a labeled real data set containing data from different variables measured at that WWTP is provided.

4.1 Experiment's setup

This section provides the experiment setups, including the tools and metrics used to measure and compare the performance of the regression methods. To set up the experiments, the fundamental stages of machine learning problems are summarized in data preprocessing, feature selection and application of regression.

4.1.1 Data preprocessing This process encompassed several operations aimed at ensuring the quality and utility of the data before conducting any analysis or modeling. Here is a more detailed explanation of the actions taken during this initial step:

- Identification of Missing Data: an analysis of the dataset was conducted to identify samples containing missing data. Missing data can arise from various reasons, such as collection errors, measurement failures or simply a lack of information. Identifying and addressing these cases is crucial to avoid biases or inaccuracies in the final results.
- Removal of Constant Variables: variables with the same value across all samples were identified, indicating that they did not contribute discriminative information to the analysis. These variables were removed from the dataset as they would not add variability and would not be useful for subsequent analysis.
- Imputation of Missing Data: for variables with missing values, an imputation method was used to estimate and fill in the missing values. In this case, the missing values were replaced with the mean value of the corresponding variable across the entire dataset. In addition, several common imputation methods, such as median-based imputation, nearest neighbor imputation and multiple imputation by regression, were considered and evaluated for suitability. Each imputation method has its advantages and limitations, and the selection of the most appropriate method was based on the characteristics of the dataset and the specific requirements of the analysis. Data imputation plays a critical role in maintaining the integrity of the dataset and ensuring that essential information is not lost during the analysis.

In summary, the first preprocessing step focused on ensuring the quality and completeness of the dataset by addressing issues such as missing data, removal of constant variables and imputation of missing values. These actions are essential to establish a solid foundation before undertaking more advanced analyses and statistical modeling.

4.1.2 Feature selection After preparing the dataset, feature selection methods were employed to examine potential correlations between the input variables and the output variable. For each method applied, a correlation value was derived for every input variable concerning total nitrogen (output variable).

Among the four feature selection methods utilized, two involve specific parameters for adjusting the degree of feature penalty (Random Forest and LASSO). These parameters influence the

TABLE 2. Tested parameters.

Reg. Tech.	Parameter	Possible settings	Description
LR	Fit Intercept (FI)	True/False	It activates or not the intercept for the model.
	Positive (P)	True/False	It forces or not the coefficients to be positive.
KNN	N_Neighbors (NN)	[2 to 20], step of 2	Number of neighbors to use.
	Weights (W)	Distance Uniform	Weight function used in prediction. Weights can be uniform, i.e. all points in each neighborhood are weighted equally. Or points can be weighted by the inverse of their distance.
DT	Criterion (C)	Squared_Error Absolute_Error	Function to measure the quality of a split. On the one hand the 'squared_error' criterion for the mean squared error. On the other hand, the 'absolute_error' criterion for the mean absolute error.
	Max Depth (MD)	[10 to 100], step of 10	The maximum depth of the tree.

dependence between variables. It is imperative to fine-tune the number of estimators, i.e. the maximum number of trees to be constructed, which, in this instance, was set at 100. Regarding LASSO, an alpha value must be adjusted, and its restrictiveness increases with higher values. In this study, an alpha value of 0.2 was chosen. These parameter settings were determined through experimental adjustments.

4.1.3 Regression After obtaining the characteristics most correlated with performance through each of the four selection methods, the next crucial step involves testing these relationships using regression techniques.

The first step in this phase is to adjust the parameters of each regression method to achieve robust and reliable results in the subsequent analysis. Parameters play a crucial role in influencing the performance and generalization capabilities of regression models. Through a process of repeated experimental fitting and cross-validation with a *k-fold* of 10, the parameters of each model are refined. In this process, the data set is divided into 10 equal parts: in each iteration, nine of these parts are used as the training set and the tenth part is used as the validation set. This procedure is repeated ten times, each time changing the validation subset, ensuring that every part of the data set is used for both training and validation. This not only allows for a thorough evaluation of model performance, but also ensures that the results are generalizable and do not depend on a single specific partition of the data.

In each repetition of the process, a model with different parameters is tested. For each of the techniques used a sweep is made with different values for the parameters to be adjusted. Table 2 shows all the possibilities for each parameter, as well as a brief explanation of their role.

4.1.4 Verification of results Once the models have been built, their accuracy and predictive power must be verified, for which different performance indicators are used:

- Mean Absolute Error (*MAE*): The average of the absolute differences between the predicted values and the actual values.
- Mean Square Error (*MSE*): The average of the squared differences between the predicted values and the actual values.
- Root Mean Square Error (*RMSE*): The square root of the average of the squared differences between the predicted values and the actual values, providing a measure of the standard deviation of the prediction errors.
- Symmetric Mean Absolute Percentage Error (*SMAPE*): A measure of the relative accuracy of the predictions, calculated as the average of the absolute differences between the predicted and actual values divided by the sum of the predicted and actual values, multiplied by 100.
- Coefficient of Determination (R^2): A measure of how well the regression model fits the observed data, indicating the proportion of the variance in the dependent variable that is predictable from the independent variables.

For all indicators, mean values are obtained to summarize the overall performance of the models.

4.2 Results

4.2.1 Data preprocessing After having processed the data, it is obtained a data set with 21 variables in total, since two of them (NO2_E, NO2_S) have been eliminated for having constant values that do not contribute to the richness of the model. The percentage of imputed data in the whole dataset is 10% of 21 variables monitored in the case study plant.

4.2.2 Feature selection To determine the correlation between the initial set of variables and the chosen output, the four characteristic selection methods discussed above are run.

LASSO The first method yields the results shown in Figure 2.

For the formation of the cluster derived from applying the LASSO method, the two variables with the highest values of relative importance were chosen, taking into account that in this methodology values close to zero are assigned to the variables that can be disregarded. It should be noted that the most influential variable, with 83% of relative importance concerning total nitrogen, presents a direct correlation. The cluster of input variables is composed of: 1) Total Kjeldahl Nitrogen at the output, and 2) Nitrate at the output.

Matrix Correlation Second method yields the results shown in Figure 3.

A threshold of 57% relative importance with total nitrogen is established, thereby configuring a cluster comprising three variables, all characterized by a direct correlation with the output. The cluster of input variables that is formed is composed of: 1) Total Kjeldahl Nitrogen at the output, 2) Nitrate at the output and 3) Total Phosphoro at the output.

Mutual Info Regression Third yields the results shown in Figure 4.

A relative importance threshold of 38% with total nitrogen is established, resulting in the formulation of a two-variable cluster, all featuring a direct correlation with the output. The cluster of input variables that is formed is composed of: 1) Total Kjeldahl Nitrogen at the output, and 2) Total Phosphoro at the output.

Random Forest Fourth method yields the results shown in Figure 5.

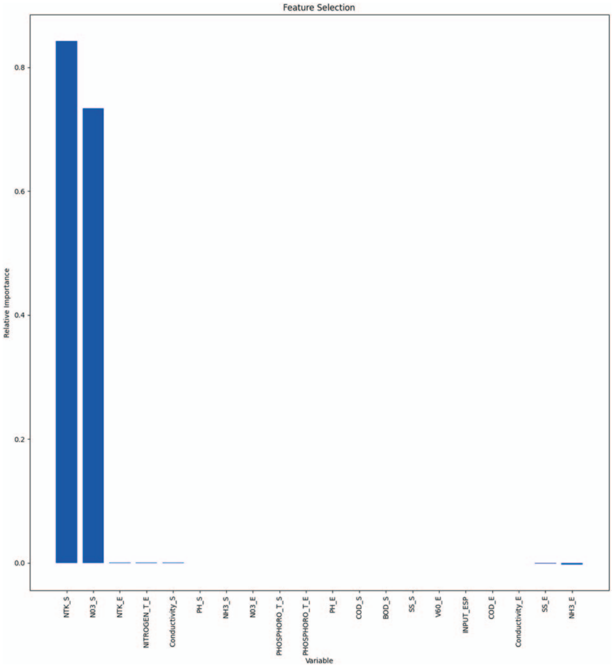


FIGURE 2. Results of feature selection methods for LASSO.

A relative importance threshold of 25% with total nitrogen is instituted, giving rise to a two-variable cluster, both exhibiting a direct correlation with the output. The cluster of input variables that is formed is composed of: 1) Total Kjeldahl Nitrogen at the output, and 2) Nitrate at the output.

4.2.3 Regression Utilizing the identified subgroups of correlated variables, the selected regression techniques are individually applied to each of the four subgroups. The iterative approach used ensures that the parameters chosen are not arbitrary, but are optimized for the specific characteristics of the data set and the objectives of the analysis. The adjusted parameters for each model, distinguished by the four variable selection methods, are compiled and presented in Table 3.

This approach ensures that each subgroup is subjected to an optimized regression analysis tailored to its specific characteristics. A summary of the calculated metrics is presented in Table 4. This table contains the chosen performance metrics: MSE, MAE, $RMSE$, $SMAPE$ and R^2. These metrics serve as quantitative measures to assess the accuracy and predictive power of the regression models for each subgroup.

Following the analysis of the metrics in Table 4, the most appropriate model is selected for each of the variable clusters. The LR model emerges as the optimal choice for the clusters generated through the LASSO, Correlation Matrix and Random Forest methods. Conversely, the KNN model demonstrates superior efficacy for the cluster arising from the Mutual Info Regressions method.

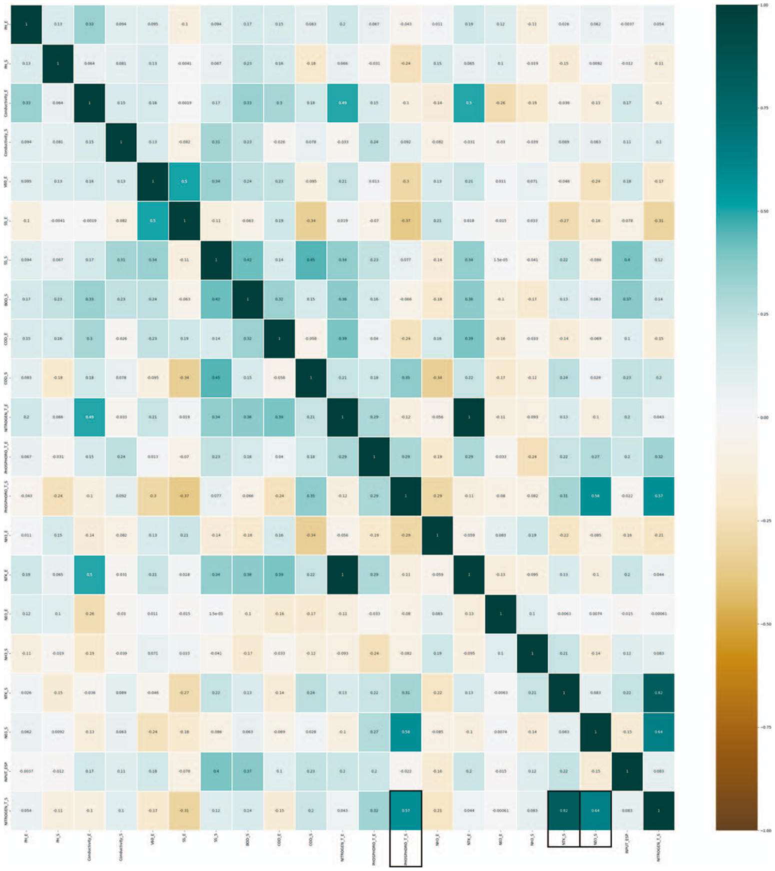


FIGURE 3. Results of feature selection methods for Correlation Matrix.

With the best model selected, the merit of these is evaluated by predicting the total nitrogen for the test data set (16% of the total dataset). The results derived from the predictions of the four models, each corresponding to a different set of variables, are plotted in Figure 6.

Overall, each of the four models demonstrates a commendable ability to accurately and acceptably predict the output. Some instances reveal the identification of outliers, typically manifesting as minimum values. Notably, the LR model emerges as the most adept in terms of fitting and prediction, underscoring its superior performance in three of the four clusters of variables selected.

For a numerical justification of the graphical results obtained in Figure 6, we provide a detailed comparison of the predictions generated by the different chosen models. Table 5 presents the best overall results, not only for feature selection, but also for each of the models used in the analysis. These graphical results illustrate the accuracy and predictive power of each model, allowing a visual assessment of its performance. The combination of the numerical and graphical results provides a complete understanding of the performance of the models and supports the conclusions derived from the analysis.

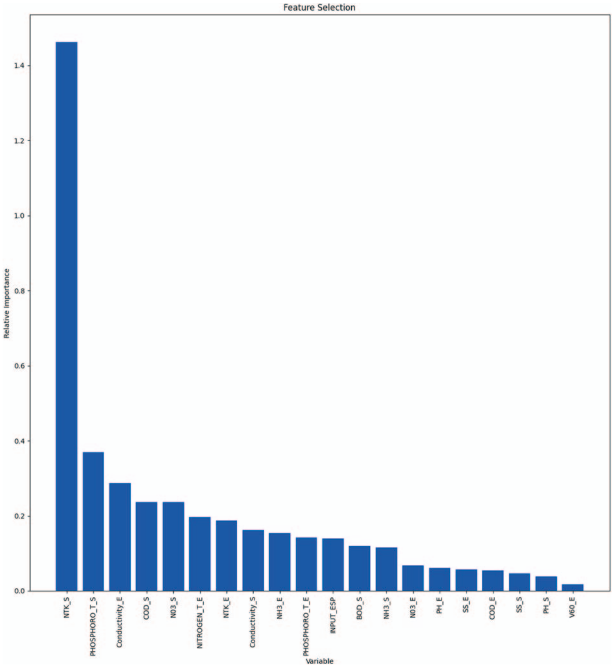


FIGURE 4. Results of feature selection methods for Mutual Info Regression.

TABLE 3. Parameter setting for regression techniques.

Regression Model	LASSO	Correlation Matrix	Mutual Info Reg.	Random Forest

LR	FI = True
	P = True
		FI = True
		P = True
			FI = True
			P = True
				FI = True
				P = True

KNN	NN = 10
	W = distance
		NN = 2
		W = distance
			NN = 10
			W = distance
				NN = 2
				W = distance

DT	C = squared
	MD = 100
		C = squared
		MD = 10
			C = squared
			MD = 100
				C = squared
				MD = 10

5 Conclusions and future work

This research delves into the analysis of feature selection methods aimed at identifying the input variables most strongly correlated with the output variable. Subsequently, these selected variables are tested using regression techniques to obtain a model capable of predicting the chosen output variable with a cluster of a few inputs.

The conclusions obtained from this study highlight the potential for generalizing the model using the LR method. Specifically, when predicting total nitrogen, it is noteworthy that the feature selection across four different techniques is very similar. This consistency, focusing on two to three variables, implies substantial savings in terms of input sensors.

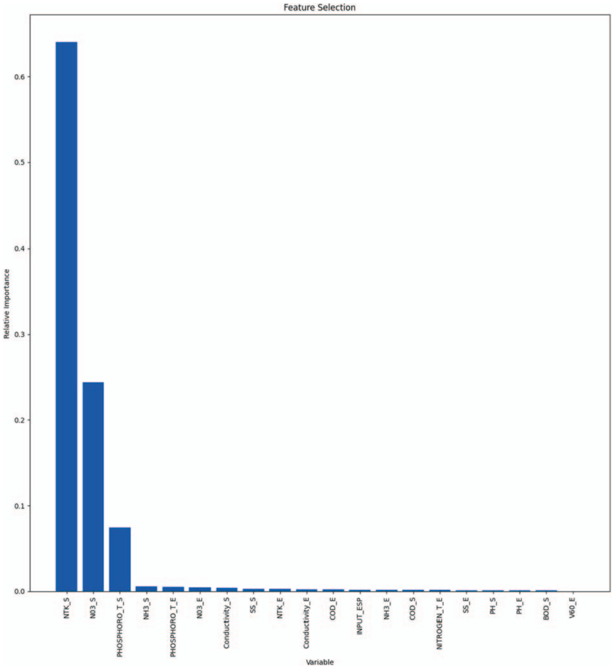
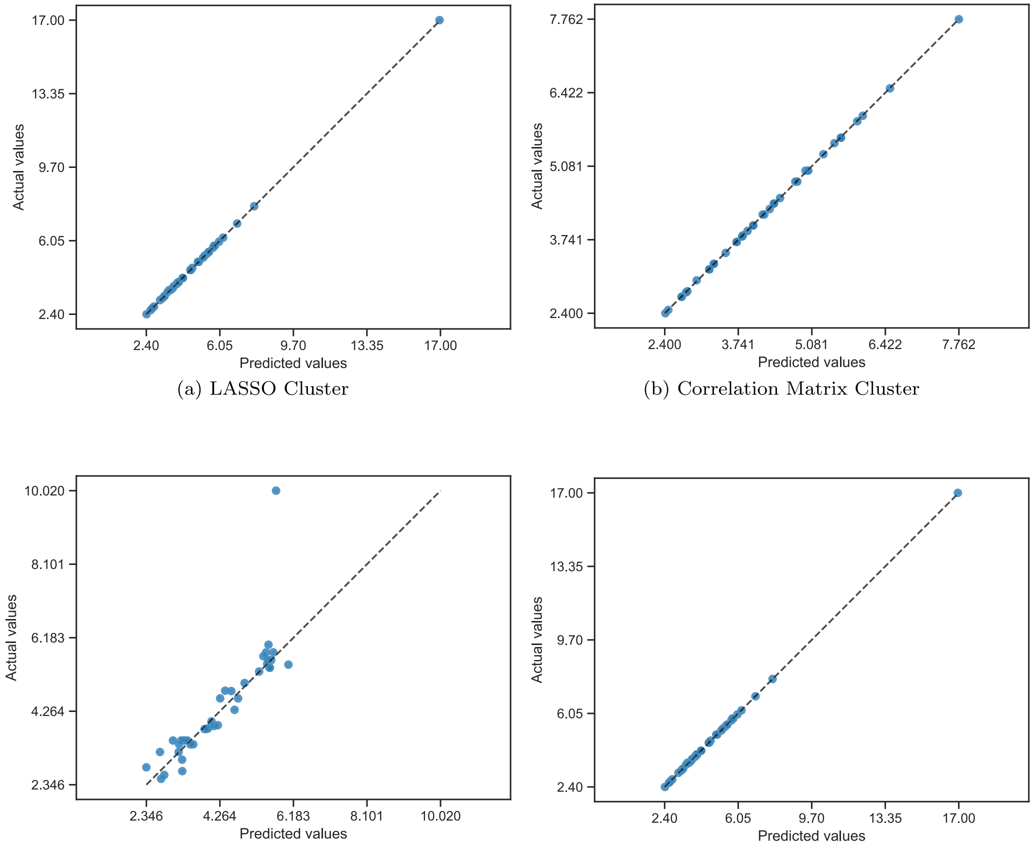


FIGURE 5. Results of feature selection methods for Random Forest.

TABLE 4. Metrics of regression models as a function of feature selection methods.

Feat. Select. Method	Reg. Tech.	MAE	MSE	RMSE	SMAPE	R^2
LASSO	**LR**	**0.01366**	**0.00035**	**0.01835**	**0.17083**	**0.99972**
	KNN	0.14526	0.45423	0.42266	1.24452	0.90989
	DT	0.18383	0.45132	0.49635	15.99336	0.83981
Correlation Matrix	**LR**	**0.01378**	**0.00034**	**0.01816**	**0.17062**	**0.99978**
	KNN	0.18445	0.32657	0.41938	1.87122	0.91477
	DT	0.20673	0.43478	0.50821	16.28372	0.86444
Mutual Info Reg.	LR	0.43304	0.54852	0.66104	5.08703	0.77209
	KNN	**0.36317**	**0.80643**	**0.68150**	**3.93581**	**0.77714**
	DT	0.40308	0.69996	0.71729	16.01914	0.72456
Random Forest	**LR**	**0.01366**	**0.00035**	**0.01835**	**0.17083**	**0.99972**
	KNN	0.14526	0.45423	0.42266	1.24455	0.90989
	DT	0.18383	0.45132	0.49635	15.9933	0.83981

It has been determined that, of the 23 variables monitored at the plant, only two to three are necessary to achieve optimal total nitrogen prediction. This simplification not only enhances computational efficiency but also has practical implications, allowing the extrapolation of a simple and generalizable prediction model for total nitrogen, applicable to a wide variety of WWTP. The



(a) LASSO Cluster

(b) Correlation Matrix Cluster

(c) Mutual Info Regression Cluster

(d) Random Forest Cluster

FIGURE 6. Prediction of total nitrogen for the four clusters.

TABLE 5. Metrics of test phase.

Feat. Select. Method	Reg. Tech.	MAE	MSE	RMSE	SMAPE	R^2
LASSO	LR	0.016068	0.000435	0.020862	0.215718	0.999475
Correlation Matrix	**LR**	**0.013382**	**0.000204**	**0.014296**	**0.199258**	**0.999566**
Mutual Info Reg.	KNN	0.285958	0.116877	0.341873	3.697416	0.856151
Random Forest	LR	0.016068	0.000435	0.020862	0.215718	0.999475

primary objective of this approach is to reduce the costs associated with computation and detection. Additionally, implementing this model offers two main advantages: first, the elimination of a sensor, thereby reducing related costs such as maintenance; and second, the use of predicted measurements to verify the accuracy of actual measurements, thus allowing for validation of the sensor's proper operation. Overall, these conclusions highlight the feasibility and effectiveness of the proposal, contributing to the optimization of resources and processes in the monitoring and control of water quality in treatment plants.

Future research presents an intriguing opportunity to generalize one of the established linear models, whereby the current model could be tested with different datasets. Moreover, a specific subset of variables correlated with the parameter under investigation or any other parameter of interest measured at the WWTP could be considered, and a robust selection methodology could be created. This strategic approach aims to pave the way for the implementation of virtual sensors, ensuring predictably robust performance. As an alternative application, the model could function as a systematic tool to validate the measurements obtained from a sensor responsible for measuring the selected output variable, thereby increasing the reliability of the prediction process and facilitating the operation for plant workers [15, 44].

Acknowledgements

Míriam Timiraos' research was supported by the 'Xunta de Galicia' through grants to industrial PhD (http://gain.xunta.gal/), under the 'Doutoramento Industrial 2022' grant with reference: 04_IN606D_2022_2692965. CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). This research is the result of the Strategic Project 'Critical infrastructures cybersecure through intelligent modeling of attacks, vulnerabilities and increased security of their IoT devices for the water supply sector' (C061/23), as a result of the collaboration agreement signed between the National Institute of Cybersecurity (INCIBE) and the University of A Coruña. This initiative is carried out within the framework of the funds of the Recovery Plan, Transformation and Resilience Plan funds, financed by the European Union (Next Generation). Grant PID2022-137152NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] M. P. Allen. *Understanding Regression Analysis*. Springer Science & Business Media, 2004.

[2] G. E. Ayhan Demirbas and W. M. Alalayah. Sludge production from municipal wastewater treatment in sewage treatment plant. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 2017; **39**:999–1006. https://doi.org/10.1080/15567036.2017.1283551.

[3] F. Bagherzadeh, M.-J. Mehrani, M. Basirifard *et al.* Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *Journal of Water Process Engineering*, 2021; **41**:102033. https://doi.org/10.1016/j.jwpe.2021.102033.

[4] F. Bagherzadeh, M. J. Mehrani, M. Basirifard *et al.* Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *Journal of Water Process Engineering*, 2021; **41**:102033.

[5] S. Balku. Comparison between alternating aerobic–anoxic and conventional activated sludge systems. *Water Research*, 2007; **41**:2220–8. https://doi.org/10.1016/j.watres.2007.01.046.

[6] G. Bertanza, R. Boiocchi and R. Pedrazzani. Improving the quality of wastewater treatment plant monitoring by adopting proper sampling strategies and data processing criteria. *Science of The Total Environment*, 2022; **806**:150724. https://doi.org/10.1016/j.scitotenv.2021.150724.

[7] A. Boretti and L. Rosa. Reassessing the projections of the world water development report. *NPJ Clean Water*, 2019; **2**:15. https://doi.org/10.1038/s41545-019-0039-9.

[8] S. Borzooei, G. Campo, A. Cerutti *et al.* Optimization of the wastewater treatment plant: from energy saving to environmental impact mitigation. *Science of the Total Environment*, 2019; **691**:1182–9. https://doi.org/10.1016/j.scitotenv.2019.07.241.

[9] T. C. Brown, V. Mahat and J. A. Ramirez. Adaptation to future water shortages in the united states caused by population growth and climate change. *Earth's Future*, 2019; **7**:219–34. https://doi.org/10.1029/2018EF001091.

[10] J. Carballo Mato, S. G. Vázquez, J. F. Águila *et al.* Foam segmentation in wastewater treatment plants. *Water*, 2024; **16**. https://doi.org/10.3390/w16030390.

[11] M. L. Christensen, K. Keiding, P. H. Nielsen *et al.* Dewatering in biological wastewater treatment: a review. *Water Research*, 2015. Special Issue on Sludge Research; **82**:14–24. https://doi.org/10.1016/j.watres.2015.04.019.

[12] D. L. Cunha, A. S. A. da Silva, R. Coutinho *et al.* Optimization of ozonation process to remove psychoactive drugs from two municipal wastewater treatment plants. *Water, Air, & Soil Pollution*, 2022; **233**:67.

[13] V. Fonti and E. Belitser. Feature selection using lasso, 2017.

[14] R. J. Freund, W. J. Wilson and P. Sa. *Regression Analysis*. Elsevier, 2006.

[15] J. M. Gonzalez-Cava, R. Arnay, J. A. Mendez-Perez *et al.* Machine learning techniques for computer-based decision systems in the operating theatre: application to analgesia delivery. *Logic Journal of the IGPL*, 2020; **29**:236–50. https://doi.org/10.1093/jigpal/jzaa049.

[16] J. Ianes, B. Cantoni, E. U. Remigi *et al.* A stochastic approach for assessing the chronic environmental risk generated by wet-weather events from integrated urban wastewater systems. *Environ. Sci. Water Res. Technol.*, 2023; **9**:3174–90.

[17] A. I. Ivanov, A. V. Bezyayev and A. I. Gazin. Simplification of statistical description of quantum entanglement of multidimensional biometric data using symmetrization of paired correlation matrices. *Journal of Computational and Engineering Mathematics*, 2017; **4**:3–13. https://doi.org/10.14529/jcem170201.

[18] A. Jović, K. Brkić and N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015—Proceedings*. 1200–5, 2015.

[19] A. Kizgin, D. Schmidt, A. Joss *et al.* Application of biological early warning systems in wastewater treatment plants: introducing a promising approach to monitor changing wastewater composition. *Journal of Environmental Management*, 2023; **347**:119001. https://doi.org/10.1016/j.jenvman.2023.119001.

[20] A. Kraskov, H. Stögbauer and P. Grassberger. Estimating mutual information. *Physical Review E*, 2004; **69**:066138. https://doi.org/10.1103/PhysRevE.69.066138.

[21] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja *et al.* Random forest for big data classification in the internet of things using optimal features. *International Journal of Machine Learning and Cybernetics*, 2019; **10**:2609–18. https://doi.org/10.1007/s13042-018-00916-z.

[22] H. Liu, E. R. Dougherty, J. G. Dy *et al.* Evolving feature selection. *IEEE Intelligent Systems*, 2005; **20**:64–76. https://doi.org/10.1109/MIS.2005.105.

[23] H. Liu and Y. Lei. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2005; **17**:491–502. https://doi.org/10.1109/TKDE.2005.66.

[24] S. Longo, B. M. d'Antoni, M. Bongards *et al.* Monitoring and diagnosis of energy consumption

in wastewater treatment plants. A state of the art and proposals for improvement. *Applied Energy*, 2016; **179**:1251–68. https://doi.org/10.1016/j.apenergy.2016.07.043.

[25] L. Jia-Yuan, X.-M. Wang, H.-Q. Liu *et al.* Optimizing operation of municipal wastewater treatment plants in China: the remaining barriers and future implications. *Environment International*, 2019; **129**:273–8.

[26] R Martínez, N Vela, A. el Aatik *et al.* On the use of an iot integrated system for water quality monitoring and management in wastewater treatment plants. *Water*, 2020; **12**. https://doi.org/10.3390/w12041096.

[27] F. Mascher, W. Mascher, F. Pichler-Semmelrock *et al.* Impact of combined sewer overflow on wastewater treatment and microbiological quality of rivers for recreation. *Water*, 2017; **9**. https://doi.org/10.3390/w9110906.

[28] V. Matamoros and V. Salvadó. Evaluation of a coagulation/flocculation-lamellar clarifier and filtration-uv-chlorination reactor for removing emerging contaminants at full-scale wastewater treatment plants in spain. *Journal of Environmental Management*, 2013; **117**:96–102. https://doi.org/10.1016/j.jenvman.2012.12.021.

[29] X. Mestre and P. Vallet. Correlation tests and linear spectral statistics of the sample correlation matrix. *IEEE Transactions on Information Theory*, 2017; **63**:4585–618. https://doi.org/10.1109/TIT.2017.2689780.

[30] F. Modaresi, S. Araghinejad and K. Ebrahimi. A comparative assessment of artificial neural network, generalized regression neural network, least-square support vector regression, and k-nearest neighbor regression for monthly streamflow forecasting in linear and nonlinear conditions. *Water Resources Management*, 2018; **32**:243–58. https://doi.org/10.1007/s11269-017-1807-2.

[31] R. Muoio, L. Palli, I. Ducci *et al.* Optimization of a large industrial wastewater treatment plant using a modeling approach: a case study. *Journal of Environmental Management*, 2019; **249**:109436. https://doi.org/10.1016/j.jenvman.2019.109436.

[32] R. Muthukrishnan and R. Rohini. Lasso: a feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*. 18–20, 2017.

[33] Oakley. Preliminary treatment and primary sedimentation. *Global Water Pathogen Project*, 2021.

[34] D. Orhon. Evolution of the activated sludge process: the first 50 years. *Journal of Chemical Technology; Biotechnology*, 2014; **90**:608–40. https://doi.org/10.1002/jctb.4565.

[35] M. Patziger, H. Kainz, M. Hunze *et al.* Influence of secondary settling tank performance on suspended solids mass balance in activated sludge systems. *Water Research*, 2012; **46**:2415–24. https://doi.org/10.1016/j.watres.2012.02.007.

[36] S. Porras, E. Jove, B. Baruque *et al.* A comparative analysis of intelligent techniques to predict energy generated by a small wind turbine from atmospheric variables. *Logic Journal of the IGPL*, 2022; **31**:648–63. https://doi.org/10.1093/jigpal/jzac031.

[37] J. Ranstam and J. A. Cook. Lasso regression. *British Journal of Surgery*, 2018; **105**:1348.

[38] M. Razif, B. Y. Soemarno, A. Rachmansyah *et al.* Implementation of regression linear method to predict WWTP cost for EIA: case study of ten malls in Surabaya city. *Procedia Environmental Sciences*, 2015. The 5th Sustainable Future for Human Security (SustaiN 2014); **28**:158–65. https://doi.org/10.1016/j.proenv.2015.07.022.

[39] S. Revollar, R. Vilanova, P. Vega *et al.* Wastewater treatment plant operation: simple control schemes with a holistic perspective. *Sustainability*, 2020; **12**. https://doi.org/10.3390/su12030768.

[40] B. C. Ross. Mutual information between discrete and continuous data sets. *PLoS One*, 2014; **9**:e87357. https://doi.org/10.1371/journal.pone.0087357.

[41] Y. Saeys, I. Inza and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007; **23**:2507–17. https://doi.org/10.1093/bioinformatics/btm344.

[42] H. Safarpour, M. Tabesh and S. A. Shahangian. Environmental assessment of a wastewater system under water demand management policies. *Water Resources Management*, 2022; **36**:2061–77. https://doi.org/10.1007/s11269-022-03129-w.

[43] R. Şenol, O. Salman and Z. Kaya. Potable water production from ambient moisture. *Applied Water Science*, 2023; **13**:10. https://doi.org/10.1007/s13201-022-01814-0.

[44] S. Simić, Z. Banković, J. R. Villar *et al.* A hybrid fuzzy clustering approach for diagnosing primary headache disorder. *Logic Journal of the IGPL*, 2020; **29**:220–35. https://doi.org/10.1093/jigpal/jzaa048.

[45] F. R. Spellman. *Handbook of Water and Wastewater Treatment Plant Operations*. CRC Press, 2013.

[46] S. Xiaogang, X. Yan and C.-L. Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2012; **4**:275–94. https://doi.org/10.1002/wics.1198.

[47] N. D. Vanli and S. S. Kozat. A comprehensive approach to universal piecewise nonlinear regression based on trees. *IEEE Transactions on Signal Processing*, 2014; **62**:5471–86. https://doi.org/10.1109/TSP.2014.2349882.

[48] N. D. Vanli, M. O. Sayin, M. N. Mohammadreza *et al.* Nonlinear regression via incremental decision trees. *Pattern Recognition*, 2019; **86**:1–13. https://doi.org/10.1016/j.patcog.2018.08.014.

[49] T. Windeatt. Accuracy/diversity and ensemble mlp classifier design. *IEEE Transactions on Neural Networks*, 2006; **17**:1194–211. https://doi.org/10.1109/TNN.2006.875979.

[50] G. Ye, J. Wan, Z. Deng *et al.* Prediction of effluent total nitrogen and energy consumption in wastewater treatment plants: Bayesian optimization machine learning methods. *Bioresource Technology*, 2024; **395**:130361. https://doi.org/10.1016/j.biortech.2024.130361.

[51] D. P. Zagklis and G. Bampos. Tertiary wastewater treatment technologies: a review of technical, economic, and life cycle aspects. *Processes*, 2022; **10**. https://doi.org/10.3390/pr10112304.

[52] F. Zayas-Gato, E. Jove, J.-L. Casteleiro-Roca *et al.* Intelligent model for active power prediction of a small wind turbine. *Logic Journal of the IGPL*, 2022; **31**:785–803. https://doi.org/10.1093/jigpal/jzac040.

[53] S. Zhang, X. Li, M. Zong *et al.* Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology*, 2017; **8**:1–19. https://doi.org/10.1145/2990508.

Received 2 August 2024