UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

LETÍCIA RODRIGUES NERI

FEASIBILITY ASSESSMENT OF DATA-DRIVEN MODELING FOR A WASTEWATER TREATMENT PROCESS IN CEARÁ

FORTALEZA

2025

LETÍCIA RODRIGUES NERI

FEASIBILITY ASSESSMENT OF DATA-DRIVEN MODELING FOR A WASTEWATER TREATMENT PROCESS IN CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Orientadora: Prof. Dr. Michela Mulas

Coorientador: Eng. Me. Geraldo Jorge Damasceno de Medeiros

FORTALEZA

2025

LETÍCIA RODRIGUES NERI

FEASIBILITY ASSESSMENT OF DATA-DRIVEN MODELING FOR A WASTEWATER TREATMENT PROCESS IN CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Michela Mulas (Orientadora)
Universidade Federal do Ceará (UFC)

Eng. Me. Geraldo Jorge Damasceno de Medeiros (Coorientador)
Companhia de Água e Esgoto do Ceará (CAGECE)

Prof. Dr. George Pereira Thé
Universidade Federal do Ceará (UFC)

"Errors using inadequate data are much less than those using no data at all."

(Charles Babbage)

ABSTRACT

The optimization of wastewater treatment plant (WWTP) operations is increasingly essential in the context of growing environmental demands, industrial variability, and regulatory tightening, particularly following the implementation of Brazil's New Sanitation Legal Framework (Law No. 14,026/2020). Among the available treatment technologies, activated sludge processes stand out for their operational flexibility and efficiency in removing organic matter and nutrients. However, the inherent complexity and dynamic behavior of these systems pose significant challenges for real-time monitoring and control.

In response to these challenges, this study proposes a data-driven modeling framework for an industrial activated sludge WWTP, aimed at predicting key performance indicators such as dissolved oxygen and sulfate concentration. The research utilizes real operational data and implements a suite of time series forecasting techniques, including ARIMA, SARIMAX, and multivariate regression models, combined with cross-validation strategies suitable for temporal data structures.

A notable aspect of this study is the careful consideration of data quality limitations, which emerged due to inconsistencies and missing values naturally associated with the operational team's decision-making. Rather than excluding these challenges, the study incorporates them as intrinsic to real-world industrial contexts and evaluates their impact on model performance.

The results demonstrate the feasibility of applying data-driven models to enhance process observability and support operational decision-making in WWTPs, while also underscoring the importance of investing in data governance, operator training, and the gradual development of a data-driven operational culture. The study concludes with recommendations for improving data management practices and advancing the analytical maturity of wastewater treatment operations in line with national regulatory requirements and international best practices.

Keywords: Wastewater treatment. Process modelling. Temporal series. Data analysis.

RESUMO

A otimização da operação de Estações de Tratamento de Efluentes (ETE) tornou-se uma demanda crescente diante do aumento das exigências ambientais, da variabilidade de cargas industriais e do endurecimento regulatório, especialmente após a promulgação do Novo Marco Legal do Saneamento no Brasil (Lei nº 14.026/2020). Dentre as tecnologias disponíveis, os processos de lodos ativados destacam-se pela eficiência na remoção de matéria orgânica e nutrientes, mas apresentam elevada complexidade dinâmica, o que dificulta o monitoramento e o controle em tempo real.

Como resposta a esses desafios, este estudo sugere uma estrutura de modelagem orientada por dados para uma ETE industrial operando com processo de lodos ativados, com o objetivo de prever indicadores-chave de desempenho, como a concentração de oxigênio dissolvido e de sulfato. A pesquisa utiliza dados operacionais reais e implementa um conjunto de técnicas de previsão de séries temporais, incluindo modelos ARIMA, SARIMAX e regressões multivariadas, integradas a estratégias de validação cruzada adequadas à estrutura temporal dos dados.

Um aspecto relevante deste trabalho é a consideração criteriosa das limitações associadas à qualidade dos dados, decorrentes de inconsistências e lacunas naturalmente associadas ao processo decisório da equipe operacional. Em vez de excluir essas dificuldades, o estudo as incorpora como características inerentes ao contexto industrial real e avalia seu impacto sobre o desempenho dos modelos.

Os resultados demonstram a viabilidade da aplicação de modelos orientados por dados para ampliar a observabilidade do processo e subsidiar a tomada de decisão operacional em ETEs, ao mesmo tempo em que ressaltam a importância de investimentos em governança de dados, capacitação dos operadores e no desenvolvimento gradual de uma cultura operacional orientada por dados. O estudo é concluído com recomendações voltadas ao aprimoramento das práticas de gestão da informação e ao avanço da maturidade analítica das operações de tratamento de efluentes, em consonância com as exigências regulatórias nacionais e com as melhores práticas internacionais.

Palavras-chave: Tratamento de efluentes. Modelagem de processos. Séries temporais. Análise de dados.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AR	AutoRegressive
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
ASM	Activated Sludge Model
BOD	Biological Oxygen Demand
COD	Chemical Oxygen Demand
DO	Dissolved Oxygen
IWA	International Water Association
LASSO	Least Absolute Shrinkage and Selection Operator
LOCF	Last Observation Carried Forward
MA	Moving Average
MAE	Mean Absolute Average
OLS	Ordinary Least Squares
P and I	Piping and Instrumentation
PACF	Partial Autocorrelation Function
RMSE	Root Mean Square Error
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors
VAR	Vector AutoRegression
WWTP	Wastewater Treatment Plant

CONTENTS

1 INTRODUCTION

1.1 Thesis motivation

In recent times, environmental concerns and increasing regulations have driven the evolution of wastewater treatment systems worldwide. Specifically, in Brazil, this movement has been strengthened by the approval of the New Sanitation Legal Framework (Law No. 14,026/2020), which establishes ambitious targets for the universalization of water and wastewater services, along with stricter standards for effluent quality and operational performance.

Traditionally, decision-making in wastewater treatment plants WWTP has relied on expert knowledge, empirical correlations, and deterministic models. Among the methods, activated sludge plants stand out for their operational flexibility and efficiency in removing organic matter and nutrients. (ORHON *et al.*, 2009). However, the complexity of their biological and physicochemical processes, along with the variability in influent characteristics, especially in plants treating industrial wastewater, poses significant challenges to process monitoring, control, and optimization. Despite their widespread application, activated sludge systems are highly dynamic, nonlinear, and sensitive to variations in influent composition, environmental conditions, and operational parameters. This complexity presents significant challenges for process monitoring and control, particularly in industrial WWTPs where influent characteristics can fluctuate substantially.(AFAN *et al.*, 2024)

In parallel, the increasing availability of operational data, combined with advances in data science and artificial intelligence, has enabled the development of data-driven models capable of capturing complex relationships between operational variables and treatment performance indicators. Data-driven approaches, which rely on empirical data rather than explicit mechanistic formulations, have been successfully applied for fault detection, performance forecasting, and advanced process control in WWTPs (BAHRAMIAN *et al.*, 2023). These methods offer the potential to enhance process observability, improve operational efficiency, and support proactive decision-making.

Nevertheless, the effectiveness of data-driven models is highly dependent on the quality, consistency, and completeness of the data available. In practice, data collected in WWTPs often exhibit issues such as missing values, sensor failures, and operational inconsistencies, which can impair model reliability and predictive accuracy (NEWHART *et al.*, 2019).

1.2 Objectives

This work aims to assess the potential of data-driven models to support operational decisions and to critically analyze the impacts of data quality on model performance, suggesting recommendations for improving the data management practices within the plant. For the task, the following objectives are defined:

- To characterize the operational data set, identifying patterns, inconsistencies, and opportunities for process observability improvement.
- To apply and compare the performances of various time series forecasting techniques, including ARIMA, SARIMA and multiple regression models, for predicting oxygen and sulfate concentrations using operational data from a fullscale WWTP.
- To provide recommendations for improving data management practices, operator training and the use of advanced analytics to align with national regulatory regulaments and international practice.

1.3 Thesis organization

Following the introduction to the work hereby presented, Chapter 2 outlines the theoretical fundamentals necessary for understanding the modeling of wastewater treatment plants (WWTPs), focusing on activated sludge systems and data-driven modeling techniques, including regression and time series analysis. Chapter 3 describes the materials and methods used in the study, including the case study plant, data acquisition procedures, preprocessing techniques, and model development strategy. Chapter 4 presents the results obtained from the exploratory data analysis and model performance evaluations, highlighting key findings in dissolved oxygen and sulfate concentration prediction. Finally, Chapter 5 offers concluding remarks, discusses the implications of the findings, and suggests directions for future research in improving data governance and analytical practices in WWTP operations.

2 THEORETICAL FUNDAMENTALS

The optimization of WWTP operations has become a central concern in light of increasing environmental constraints, energy efficiency targets, and legal requirements such as Brazil's New Sanitation Legal Framework (Law No. 14,026/2020). Among biological treatment systems, the activated sludge process is particularly significant for its effectiveness, flexibility and ability to handle variable loads of organic and inorganic compounds. (**??**) However, it is also marked by intrinsic complexities such as strong nonlinearity and time-varying dynamics, which challenge traditional control and monitoring approaches. These complexities motivate the integration of data-driven models and advanced analytics as complementary tools for observability, forecasting, and operational support.

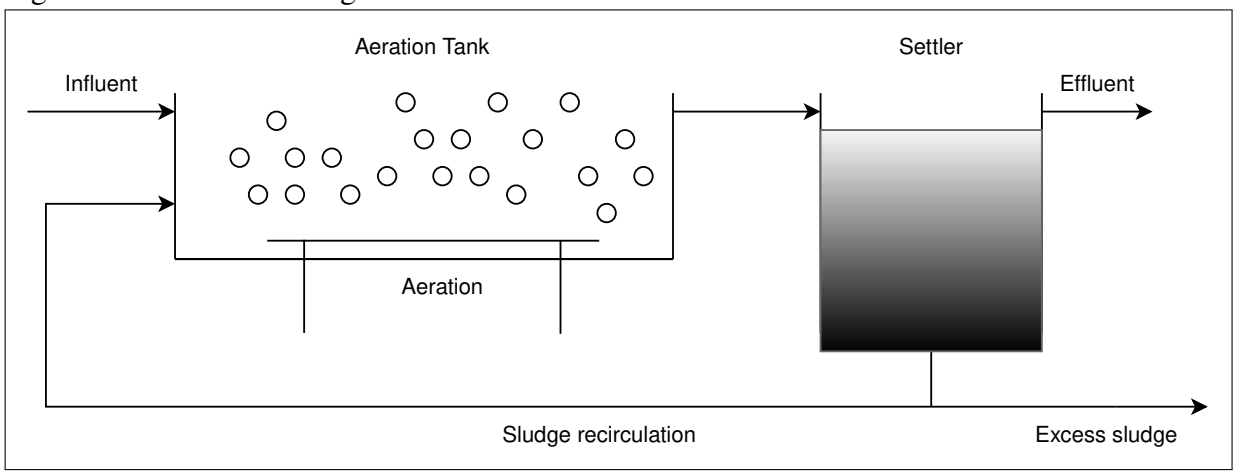
This chapter reviews the current state of the art on data-driven modelling in WWTPs, particularly within industrial systems by exploring the evolution of soft-sensors, time series forecasting approaches, and predictive control strategies, with a focus on the implications of data quality limitations for model reliability and deployment, while also reviewing the theoretical basis needed to present the methods applied during this study.

2.1 Activated Sludge Processes in Industrial Wastewater Treatment

Activated sludge systems are widely applied for both municipal and industrial effluents due to their proven effectiveness in removing biodegradable organic matter and nutrients. The fundamental principle behind the process is the cultivation and sustenance of a diverse, aerobic microbial community within a controlled environment 1. Wastewater is introduced into an aeration tank, where it is mixed with recycled biomass (activated sludge) and oxygen is supplied, fostering the oxidation of organic and certain inorganic pollutants. The mixed liquid is then routed to a secondary settling tank, enabling the separation of treated effluent from the biomass, which is then largely returned to maintain active microbial populations (TCHOBANOGLOUS *et al.*, 2003).

Despite the success of activated sludge processes in municipal settings (JENKINS; WANNER, 2014), their implementation in industrial wastewater treatment still brings forth unique challenges, as industrial influents often differ greatly from municipal ones, not only in the overall loading of organic and inorganic constituents but also in variability, shock loads, the presence of inhibitory or toxic compounds, and intermittent discharge patterns linked to

Figure 1 – Activated Sludge Process



Source: made by the author

production scheduling. For instance, flows from chemical manufacturing, food processing, or pharmaceutical plants may exhibit pronounced fluctuations both daily and seasonally, sometimes accompanied by high concentrations of recalcitrant or inhibitory substances. These features necessitate the design of tailored process configurations, ranging from specialized pretreatment steps to enhanced nutrient removal or advanced oxidation, and dynamic operational control regimes (ORHON *et al.*, 2009).

Operationalizing such flexibility, however, is challenging, as sudden changes in influent quality or hydraulic load can rapidly disrupt the biological equilibrium of the aeration basin, risking loss of biomass, the onset of bulking or foaming episodes, or failures to meet regulatory discharge limits. Therefore, the continuous monitoring of key process parameters such as dissolved oxygen, suspended solids, ammonia, becomes essential (BAHRAMIAN *et al.*, 2023). Furthermore, maintaining both effluent quality and operational efficiency in the face of such variability requires responsive, often automated, management strategies that combine high-frequency data acquisition, advanced process control algorithms, and data-driven decision support systems (NEWHART *et al.*, 2019).

Mechanistic modelling frameworks such as the Activated Sludge Model (ASM) family have demonstrated significant value, particularly in well-characterized municipal environments where reaction kinetics, process configuration, influent composition, and stoichiometry are well understood (HENZE *et al.*, 2006). However, their practical utility diminishes in industrial contexts, where influent characteristics are highly variable, non-standard, or poorly documented, and operational configurations change frequently (ORHON *et al.*, 2009). The extensive site-specific data and calibration required for mechanistic models often render them prohibitively

expensive or even unfeasible in these scenarios.

2.2 Data-Driven Modelling in Wastewater Treatment

In recent years, data-driven models have emerged as valuable tools for supplementing or replacing traditional mechanistic models in WWTP process control and performance prediction. Unlike traditional mechanistic models, based on mass balance and reaction kinetics, data-driven models extract structure, patterns, and predictive relationships directly from operational or monitoring data, as opposed to being derived from first-principles, descriptions of underlying physical, chemical, or biological processes (BAHRAMIAN *et al.*, 2023).

In this context, the feasibility of using data-driven software sensors to estimate process variables traditionally monitored by hardware instrumentation has been demonstrated (HAIMI *et al.*, 2013), with results such as the reduction of maintenance costs and the improvement of process observability. In these situations, data-driven models, such as artificial neural networks, support vector machines, random forests, or multiple regression offer a pragmatic alternative. Instead of explicitly representing the internal workings of the system, they rather learn the relationships between inputs and outputs (such as influent conditions, effluent quality, sensor readings and process parameters) solely from available historical data, with minimal assumptions regarding causal structure (NEWHART *et al.*, 2019; BAHRAMIAN *et al.*, 2023). This data-centric approach is especially useful in industrial WWTPs, where operational records, online sensor data, and laboratory measurements may exist in abundance, but mechanistic understanding or influent stability is lacking.

The data-driven approach to modelling offers significant advantages in these scenarios. First, these models can be trained quickly using data that are already available, enabling rapid deployment for monitoring, anomaly detection, or forecasting, even in plants lacking comprehensive instrumentation or detailed documentation (NEWHART *et al.*, 2019). Second, they are able to capture complex, nonlinear relationships and unanticipated interactions between process variables, which may be difficult to model using explicit physical or biochemical equations. In fact, it has been shown that data-driven software sensors based on statistical learning not only improved the observability of complex industrial WWTPs but also facilitated reductions in hardware maintenance and calibration burden (DÜRRENMATT; GUJER, 2012).

Nonetheless, the limitations to these models include their lack of interpretability, which may hinder their adoption in safety-critical environments, along with overfitting to noise

or artifacts in the training data can lead to poor generalization. Additionally, the success of these models depends fundamentally on data quality, coverage, and representativeness; scarce or biased datasets will inevitably yield less reliable predictions (NEWHART *et al.*, 2019). For this reason, the current literature increasingly advocates hybrid approaches, in which the empirical strength of black box models is combined with mechanistic constraints or expert domain knowledge to enhance robustness and operational validity. Recent work by Bahramian *et al.* (2023) illustrates how interpretable data-driven models, when combined with process understanding and variable importance analysis, can enhance operational transparency and support informed decisions about sensor placement and control optimization in WWTPs.

Hybrid decision support systems that integrates fuzzy logic with deep neural networks and data augmentation strategies have been proposed by Cosenza *et al.* (2025) for this type of system, demonstrating that such hybrid systems can bridge the gap between theoretical process optimization and real-world operational constraints, especially in contexts where infrastructure limitations and regulatory compliance pressures coexist. Furthermore, Elsayed *et al.* (2022) compared a plethora of machine learning classification algorithms applied to wastewater treatment risk prediction have been compared, identify ensemble methods as providing superior performance in terms of classification accuracy and interpretability, highlighting the importance of feature selection and variable importance analysis for practical deployment in WWTP contexts, contributing to the understanding of influent characteristics' impact on effluent quality. Indeed, despite the fact that this approach offers cost-effective alternatives to extensive physical monitoring networks, Sperling *et al.* (2020) stresses that such models must be incorporated into structured monitoring programmes, with consistent data governance protocols ensuring data quality, completeness, and traceability .

In such a context, while data-driven models do not completely replace the need for mechanism-based modelling, they fill a vital role in industrial wastewater treatment scenarios where classical process knowledge is incomplete or wholly unavailable as their capacity to quickly extract meaningful predictive relationships from operational data makes them an essential tool for performance optimization, early warning, and informed process management in today's dynamic industrial WWTP landscape.

2.2.1 Data Governance in Industrial Environments

As industrial operations, including wastewater treatment plants, increasingly embrace digitization and data-driven methodologies, the role of data governance has become critical. Data governance refers to the policies, processes, standards, and controls that ensure the effective management of data across its lifecycle: from collection and storage to analysis, sharing, and deletion (KHATRI; BROWN, 2010). In industrial contexts, robust data governance not only underpins regulatory compliance but also acts as a prerequisite for the successful deployment of advanced analytics and process optimization initiatives (BATINI, 2016).

It is then clear that data acquisition and management in industrial WWTPs entail unique complexities that distinguish these environments from their municipal counterparts or other process industries. In fact, since the operational context is characterized by irregular production schedules, intermittent industrial discharges, highly variable influent compositions, and frequently only partial automation of data collection and control systems, as a consequence, the operational data are often incomplete, asynchronously sampled, and subject to a range of data quality issues, including measurement errors, sensor drift, and inconsistencies stemming from both technical and human factors (ORHON *et al.*, 2009). Newhart *et al.* (2019) emphasize that practical applications of data-driven technologies in WWTPs routinely encounter poor data quality, missing values, and irregular time-stamping, which pose significant challenges for developing reliable predictive models, advocating then for the integration of robust data preprocessing methods, including outlier detection, temporal alignment, and missing-data imputation, as prerequisites for effective statistical modelling and forecasting.

Furthermore, stakeholder commitment and investment are critical determinants of the success of data-driven initiatives in the wastewater sector. In fact, a growing body of research highlights persistent barriers that arise in the absence of adequate stakeholder involvement, such as insufficient investment in data infrastructure, resulting in fragmented or obsolete sensor networks and unreliable data acquisition systems. In fact, inadequate financial and technical resources often lead to data quality and coverage gaps, where legacy systems and deferred maintenance manifest as inaccuracies, inconsistencies, and difficulty in integrating heterogeneous data sources (MEDEIROS *et al.*, 2025). Additionally, a lack of investment in workforce development leaves operators and staff with insufficient data literacy or analytical competencies, further diminishing the effectiveness and acceptance of new data-driven approaches (BATINI, 2016).

Organizational culture also presents a significant challenge: resistance to change among operators or management, skepticism about the value of new technologies, and lack of clear incentives to participate in data governance can result in underused or abandoned analytical systems (KHATRI; BROWN, 2010). Fragmented data governance, typified by unclear data ownership, insufficient documentation, and the absence of security or privacy protocols, renders even promising analytics initiatives vulnerable to failure, issues that are particularly pronounced when attempting to scale pilot projects to full-plant or multi-site operations without adequate coordination, investment, and cross-departmental collaboration (SUN; SCANLON, 2019).

The operational reality of many industrial WWTPs, including the case study presented in this thesis, reflects these challenges. Data inconsistencies arising from real-time acquisition lapses, gaps in laboratory-based measurements, and limited flow monitoring after key process units, such as equalization tanks, demand significant methodological adaptation. Careful preprocessing, data validation, and the adoption of advanced imputation and calibration techniques are frequently necessary to achieve even baseline data quality suitable for analysis.

Despite these hurdles, the literature points to substantial benefits resulting from effective data governance and judicious investment in data-driven technologies. Many utilities report significant returns on investment within a few years, primarily due to operational efficiencies, energy savings, predictive maintenance, and reductions in regulatory penalties (ELSAYED *et al.*, 2022). Moreover, the cost of inaction, seen in higher operational costs, increased process risk, missed funding opportunities, and greater exposure to regulatory non-compliance, can easily outweigh the initial investments required to establish robust governance frameworks. In addition, non-monetary benefits, such as increased transparency, greater stakeholder trust, and enhanced reputation underscore the long-term strategic value of adopting comprehensive data governance practices.

For these reasons, data governance should be viewed as a holistic framework that encompasses the entire data lifecycle, ensuring that data quality is not an afterthought but is built into the design of systems, workflows, and organizational roles from the outset (BATINI, 2016). Embedding these principles throughout plant operations supports not just compliance and operational excellence but also enables the effective application of advanced analytics and process innovation within industrial WWTPs.

2.2.2 Physical and Operational Characterization of a Modelled WWTP

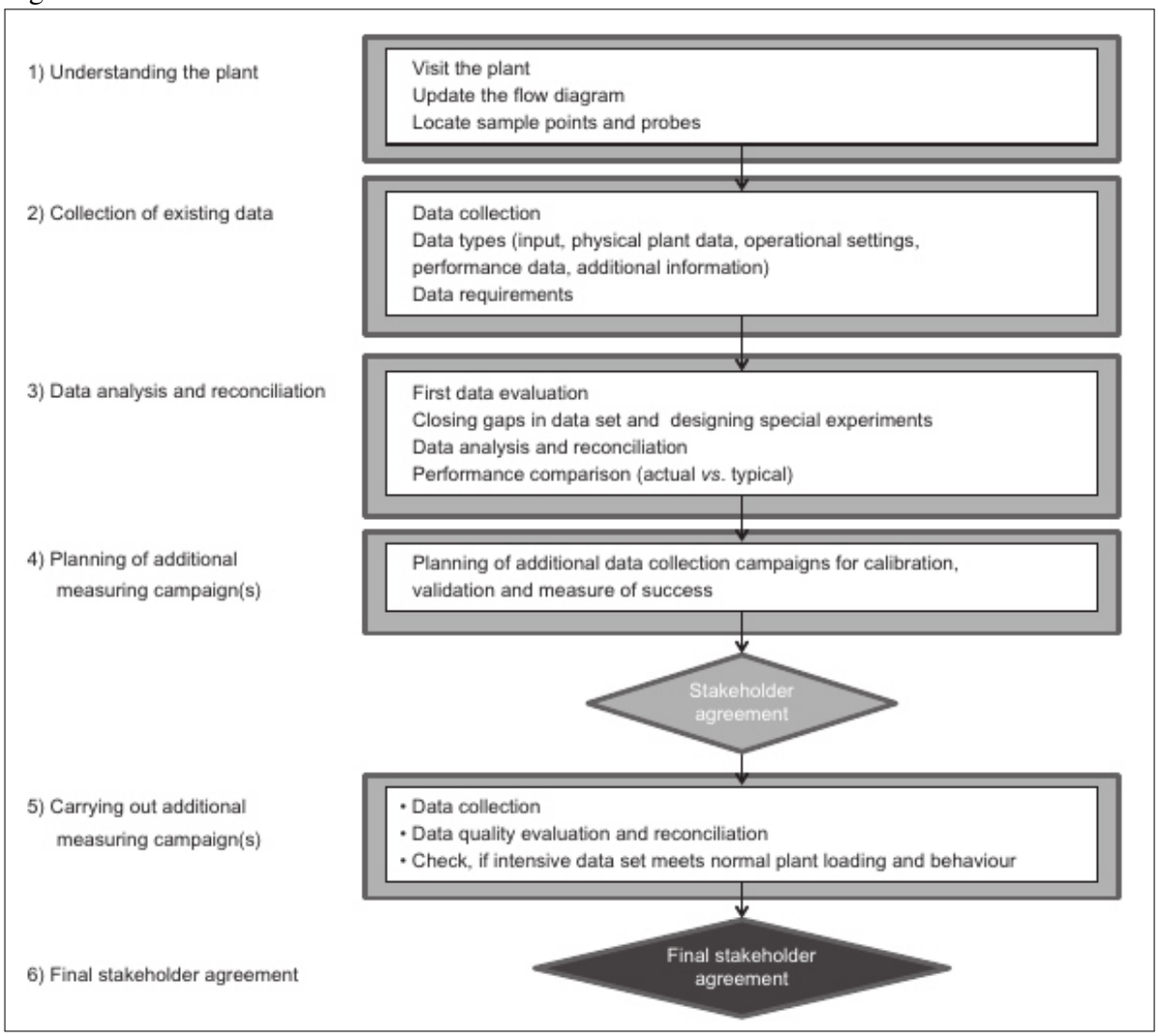
A comprehensive understanding of a WWTP relies not only on its theoretical design, but also on detailed information regarding its physical layout, operational configurations, performance metrics, and contextual factors, forming the foundation upon which reliable data-driven models and process simulations can be constructed. In fact, according to the International Water Association (IWA)'s guidelines of data collection (RIEGER *et al.*, 2012), illustrated in Figure (2), the modeling process in a WWTP begins with a comprehensive understanding of the facility, which involves reviewing documentation, analyzing the process flow diagram, conducting on-site visits, and maintaining direct communication with plant personnel. The second phase focuses on identifying available data sources, types of data, and the specific information needed to meet the modeling objectives. The following phase, consisting of data analysis and reconciliation, seeks to find inconsistencies or faults in the data sets and to produce a reliable, internally consistent database suitable for modeling. If needed, additional measurement may be conducted, though these should be approved by the plant managers due to their associated costs, and any new data collected must undergo the same quality assurance and reconciliation procedures. All data, whether existing or newly acquired, should represent typical operating conditions and must be information-rich to support calibration and forecasting objectives. Finally, the relevant stakeholders should confirm the adequacy and quality of the reconciled datasets before the model development.

This framework has been applied to our study in order to ensure a systematic and structured approach to modelling the WWTP. The process involved close collaboration with the plant personnel to validate the data, ensuring it was representative of the operating conditions. By following these steps, we were able to develop a dataset for our work.

2.2.3 Data characteristics in a WWTP

Physical plant data encompass all information related to the physical attributes and infrastructure of the WWTP, including tank dimensions (volumes, depths), their layout and hydraulic behavior, such as whether a tank operates under plug-flow or completely mixed conditions. The configuration of tanks, whether in series, in parallel, or with recycle streams, directly influences the dynamic response of the system and must be explicitly accounted for in any modelling effort. Key elements also include the locations of influent and effluent points, the

Figure 2 – Data collection and reconciliation



Source: (RIEGER et al., 2012).

type and specifications of aeration and mixing equipment (such as blowers, diffusers, valves, control schemes), and pumping infrastructure (pump capacities, operational constraints). Further, the presence and configuration of process instrumentation, sensors, actuators, and feedback control loops, should be derived from detailed Piping and Instrumentation (P&I) diagrams, as these determine the plant's ability to monitor and control key variables.

Additionally, data on sludge treatment units, including thickening, digestion, and dewatering processes, are critical. These should specify the operating regime (continuous or batch), discharge locations, and how return streams (e.g., filtrate or centrate) are managed, as they often reintroduce significant pollutant loads into the mainstream treatment line. Beyond direct operational and performance parameters, additional contextual data may offer valuable insights into plant behavior under non-standard conditions. For example, information about

upstream sewer characteristics, seasonal loading variations, industrial contributors, and chemical consumption in auxiliary units, lime in sludge dewatering can all influence system performance. (RIEGER *et al.*, 2012)

It is also crucial to document maintenance events or atypical operating conditions, such as temporary bypasses, line shutdowns, or tank cleaning procedures as informal observations from plant personnel, such as asymmetries in flow behavior or recurring anomalies during peak flows, can often help uncover hidden dynamics or inconsistencies in the recorded data. These inputs, while qualitative in nature, may provide critical validation checkpoints or hypotheses for further investigation.

2.3 Linear Regression Modelling

In a linear regression model, the goal is to model the linear relationship between a dependent variable Y and a dataset of independent variables $X_1, ..., Xp$. The fundamental assumption is that either the conditional expectation of the output variable $E(Y|X)$ is a linear function of the input variables or that the linear model serves as a plausible approximation. The input variables X_j can come from different sources, all numerical, such as quantitative inputs by themselves, transformations of quantitative inputs, such as log, square-root, squares or even basis expansions, and numeric coding of the levels of qualitative inputs (HASTIE; TIBSHIRANI, 2009).

Since linear models are mathematically simple and computationally efficient, they often provide an adequate approximation of the relationship between input and output variables and yield interpretable parameter estimates that help in understanding the effect of each predictor. Moreover, in scenarios involving small sample sizes, high noise levels, or sparse data, linear models can outperform more complex nonlinear alternatives in terms of predictive accuracy. Additionally, linear modelling techniques can be applied to transformed versions of the input variables, thereby significantly expanding their flexibility and range of applicability.

Given a dataset with n observations and p predictor variables, the linear regression model can be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{2.1}$$

where:

- $\mathbf{Y} \in \mathbb{R}^n$ is the vector of observed responses;
- $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ is the design matrix, typically including a column of ones to account for the intercept term;
- $\beta \in \mathbb{R}^{(p+1)}$ is the vector of unknown regression coefficients (including the intercept);
- $\varepsilon \in \mathbb{R}^n$ is the vector of random errors, assumed to follow a normal distribution with zero mean and constant variance, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

2.3.1 Ordinary Least Squares

Ordinary Least Squares (OLS) is the most fundamental method for estimating the parameters of a linear regression model. It is based on the principle of minimizing the sum of squared differences between the observed values and the values predicted by the linear model. These differences are known as residuals.

The goal of OLS is to find the vector $\hat{\beta}$ that minimizes the residual sum of squares (RSS):

$$\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.2}$$

The solution to this optimization problem is obtained by solving the normal equations:

$$\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{y}. \tag{2.3}$$

Provided that $\mathbf{X}^\top \mathbf{X}$ is invertible, the closed-form solution for the OLS estimator is:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2.4}$$

Among all linear unbiased estimators, OLS has the minimum variance (Gauss–Markov theorem), however, despite its simplicity and wide applicability, it has limitations in the presence of multicollinearity, high-dimensional data ($p \approx n$ or $p > n$), or when the assumption of constant variance is violated. In such cases, penalized models may provide better generalization performance or numerical stability.

2.3.2 Penalized Regression

In many practical situations, such as when the number of predictors is large or when multicollinearity exists among inputs, ordinary least squares estimation can lead to overfitting or instability in the coefficient estimates. To address these issues, regularization techniques introduce a penalty term into the loss function, effectively constraining or shrinking the coefficient estimates. This enhances both model generalization and numerical stability.

2.3.2.1 Ridge Regression

Ridge regression (also known as L_2 regularization) modifies the standard least squares loss function by adding a penalty proportional to the sum of the squared coefficients:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.5}$$

where $\lambda \geq 0$ is a tuning parameter that controls the strength of the penalty. Ridge regression shrinks the coefficients toward zero but does not set any of them exactly to zero. It is particularly useful in situations with multicollinearity, as it stabilizes the inversion of the design matrix.

2.3.2.2 LASSO Regression

Least Absolute Shrinkage and Selection Operator (LASSO) regression, or L_1 regularization, adds a penalty equal to the sum of the absolute values of the coefficients:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.6}$$

Unlike Ridge regression, the L_1 penalty has the effect of forcing some coefficient estimates to exactly zero when λ is sufficiently large. As a result, LASSO performs both regularization and variable selection, producing more interpretable models, especially in high-dimensional settings where $p \gg n$.

2.3.3 Statistical Learning and Regression Models for Industrial WWTPs

Alongside time series modelling, statistical learning and regression techniques have become increasingly central to the analysis and optimization of wastewater treatment processes. These methods enable the quantification of relationships between process variables and environmental outcomes, support root-cause analysis, and provide the foundations for predictive monitoring and process control strategies in industrial WWTPs. Their flexibility and adaptability have made statistical learning approaches especially appealing in systems characterized by complexity, multi-collinearity, and non-linear relationships. (SPERLING *et al.*, 2020)

Multiple Linear Regression (MLR), one of these methods, is still a popular modelling framework in the wastewater treatment industry, as it offers valuable insights into the dependency structures between influent characteristics, operational variables, and treatment performance metrics, despite its conceptual simplicity (DÜRRENMATT; GUJER, 2012). MLR is particularly attractive for WWTP datasets featuring lower temporal resolution, such as weekly or monthly sampling, where autocorrelation is less pronounced, and the interpretability of model coefficients is of central importance to plant operators and regulatory stakeholders.

Best practices in the deployment of regression models in industrial contexts involve rigorous adherence to statistical principles for model selection, assumption verification, and diagnostic checking. Cross-validation is commonly used to guard against overfitting and to assess the generalizability of the model to new data, a critical consideration given the inherent variability of industrial WWTP operations (NEWHART *et al.*, 2019). The detection and mitigation of multicollinearity, where predictor variables are highly correlated, are particularly important, as this phenomenon can inflate variance in coefficient estimates and diminish the reliability of inferences. (HASTIE; TIBSHIRANI, 2009)

However, traditional random-sample cross-validation techniques, commonly used in machine learning, are inappropriate for time series data because they break the temporal structure, allowing future observations to influence the model fitted to past data. To address this, time series cross-validation techniques such as rolling-origin are widely recommended. These approaches ensure that, at each validation step, the model is trained only on data available up to a certain point in time, and evaluated on future, unseen data points. Such validation closely mirrors operational forecasting scenarios in WWTPs, where models are deployed in real time to inform process control or early warning systems (HYNDMAN; ATHANASOPOULOS, 2021). In this context, preserving the temporal integrity of training and validation sets is especially crucial

for industrial settings afflicted by seasonality, shifts in operational regimes, and process upsets. Failure to do so can mask model deficiencies, particularly in the presence of autocorrelation or non-stationary dynamics typical of WWTP datasets, and can lead to process decisions based on erroneously optimistic forecasts. (NEWHART *et al.*, 2019)

More advanced statistical learning techniques, including regularized regression like Lasso and Ridge, principal components regression, and partial least squares have been adopted to further manage high-dimensional data, reduce overfitting, and enhance model interpretability in complex WWTP settings (WANG *et al.*, 2021). Additionally, non-linear and non-parametric models, such as decision trees and ensemble learning methods, are being increasingly explored for their ability to capture the intricate, non-linear relationships inherent to biological and chemical treatment processes (NEWHART *et al.*, 2019).

2.4 Time Series Modelling and Analysis

Time series are datasets in which observations are collected sequentially over time, and they appear in a wide range of real-world applications, such as the variables analysed in this work, since they all have been recorded chronologically, thus, falling under the general category of time series data.

In particular, time series are valuable for analyzing the temporal evolution of a variable or for identifying recurring patterns over similar time intervals (SHUMWAY; STOFFER, 2005). Aditionally, one of their fundamental characteristic is stochasticity, due the random or probabilistic nature of the data-generating process, which brings additional challenges for their analysis and modelling. (BOX *et al.*, 2015). Since such data generally exhibit inherent randomness, time series are formally defined as collections of random variables indexed in the order in which they were observed over time, that is, a stochastic process, where, given a sequence $x_1, x_2, x_3, \ldots$, each value x_t is assumed to have been obtained at a point in time preceding x_{t+1}. Due to these characteristics, classical regression techniques may be generally insufficient for modelling time series data, as they assume that input observations are independent from one another.

2.4.1 Autoregressive and Vector Autoregressive

This observation led to the development of AutoRegressive (AR) models, which introduce a linear dependence on past values of the series, being particularly useful for forecasting, as long as they effectively represent the dynamics of the observed data. To evaluate how well the model fits the data, statistical measures are commonly applied, providing a quantitative basis for validating the model's accuracy and reliability. AR models are typically represented by the following equation:

$$Yt = c + \varepsilon_t + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + ... + \theta_p Y_{t-p}. \tag{2.7}$$

where c is a constant term (often representing the mean of the signal), θ_i are the weighting coefficients associated with past observations, and Y_t is the model output at time tt. The term ε_t accounts for the prediction error and is typically modeled as white noise.

AR models exhibit an explicit dependence on previously observed values, and the number of past values used in the model is defined as its order, denoted by p and is typically indicated with the notation AR(p).

Once the coefficients are estimated, it becomes possible to make forecasts based on the observed data. The estimation of these coefficients can be carried out using algorithms such as the least squares method, which aims to minimize the error between the observed data and the model defined by Equation (2.7). Forecasting is performed recursively using the expression:

$$Y_{t+1} = c + \theta_1 Y_t + \theta_2 Y_{t-1} + ... + \theta_p Y_{t-p+1}. \tag{2.8}$$

This process can be repeated iteratively to generate the desired number of future predictions.

Vector AutoRegression (VAR) models are a natural extension of univariate autoregressive models to multivariate time series. While univariate models such as AR deal with single time-dependent variables, VAR models are designed to capture the linear interdependencies among multiple time series variables evolving over time.

In a VAR model of order p, denoted as VAR(p), each variable in the system is expressed as a linear combination of its own past values and the past values of all other variables in the system. This allows the model to account for both the autocorrelation within each time series and the cross-correlation between different series.

Mathematically, a VAR(p) model for a k-dimensional time series $\mathbf{y}_t = [y_{1,t}, y_{2,t}, \ldots, y_{k,t}]^T$ is defined as:

$$\mathbf{y}_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \tag{2.9}$$

where:

- $c \in \mathbb{R}^k$ is a vector of intercept terms;
- $\Phi_i \in \mathbb{R}^{k \times k}$ are coefficient matrices for lag i;
- $\varepsilon_t \in \mathbb{R}^k$ is a white noise vector with zero mean and covariance matrix Σ.

As with univariate AR models, stationarity is a fundamental requirement for VAR models. A VAR process is considered stationary if all eigenvalues of the companion matrix lie within the unit circle. If the series are not stationary, differencing or transformation methods should be applied before model fitting.

The appropriate order p of the VAR model can be determined using information criteria such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or the Hannan–Quinn Criterion (HQIC). These criteria balance model fit and complexity, helping to prevent overfitting.

The parameters of a VAR model can be estimated using multivariate least squares, since the system of equations for each variable can be treated separately but estimated jointly. Once estimated, the model can be used for dynamic forecasting of all variables in the system. The h-step ahead forecast is computed recursively by substituting future values with their predicted counterparts.

In order to evaluate stationarity in a time series, techniques such as the Augmented Dickey-Fuller (ADF) test may be applied. This test is a statistical procedure used to assess the presence of a unit root in a time series, thereby determining whether the series is stationary. A stationary series has constant mean and variance over time, while a non-stationary series may exhibit trends or other time-dependent structures. The ADF test extends the basic Dickey-Fuller test by incorporating lagged differences of the dependent variable to account for autocorrelation in the residuals. The general form of the ADF regression is given by:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t, . \tag{2.10}$$

where Δy_t is the first difference of the series, t is a time trend (optional), and ε_t is a white noise error term. The null hypothesis (H_0) of the test is that the series contains a unit root ($\gamma = 0$), implying non-stationarity, while the alternative hypothesis (H_1) is that the series is stationary ($\gamma < 0$). To evaluate this, the test statistic is compared to critical values derived from a non-standard distribution, which accounts for the presence of deterministic components (such as a constant or trend) and the sample size. These critical values are obtained through Monte Carlo simulations or asymptotic theory, as the usual t-distribution is not valid under the unit root hypothesis. If the test statistic is more negative than the critical value at a given significance level, the null hypothesis can be rejected, indicating evidence in favor of stationarity. The p-value, in this context, represents the smallest level of significance at which the null hypothesis would be rejected, providing a direct measure of the strength of evidence against non-stationarity. (MACKINNON, 1996)

Another important tool for analyzing the dynamic relationships within a VAR system is the Granger causality test. This test helps determine whether the past values of one variable X_t contain useful information for predicting another variable Y_t, beyond what is already explained by Y_t's own history. While correlation between variables does not imply causation, the temporal precedence in time series allows for a specific type of inference called predictive causality. If changes in X consistently precede and improve the prediction of changes in Y, then X is said to *Granger-cause Y* (GUJARATI; PORTER, 2009). Additionally, the test is sensitive to lag order selection; insufficient lags can miss existing causality, while excessive lags can reduce test power.

To determine whether X_t *Granger-causes* Y_t, we compare two autoregressive models: one using only past values of Y_t, and another augmented with past values of X_t. The null hypothesis is that X_t does not Granger-cause Y_t, i.e., X_t the lags of do not improve the prediction of Y_t. This hypothesis is tested via an F-test that compares the residual sums of squares (RSS) from the restricted model (without X_t) and the unrestricted model (with X_t). A significant reduction in RSS in the unrestricted model leads to rejection of the null hypothesis, suggesting that X_t provides predictive information about Y_t.

While VAR models are powerful tools for multivariate time series analysis, they assume that all variables in the system are treated symmetrically and require a large number of parameters, especially for high-dimensional data or large lag orders. This can lead to overparameterization and poor generalization if not handled carefully. Regularization techniques or

dimension reduction methods, such as Principal Component Analysis, are sometimes employed to mitigate this issue.

2.4.2 Moving Average

In contrast to AR models, Moving Average (MA) models aim to represent a random process as a function of past prediction errors. Mathematically, an Moving Average (MA) model is described by:

$$Y_t = c + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ... + \phi_q \varepsilon_{t-q}. \tag{2.11}$$

where ϕ_i are the coefficients associated with past forecast errors, and q denotes the order of the model.

The term moving average comes from the interpretation that the forecasted value Y_t is a weighted average of past prediction errors. As the forecasting process advances, these weights move along with each new prediction, continuously adjusting based on previous residuals. Similarly to AR models, the order q determines how many past errors influence the current prediction, and MA models are denoted as MA(q).

However, MA models present additional challenges during parameter estimation, as the error terms ε_t are not directly observable, which prevents the use of standard regression techniques. Instead, numerical methods are employed to estimate both the model parameters and the unobserved errors simultaneously, with the objective of fitting the model to the observed dataset.

2.4.3 Autoregressive Integrated Moving Average

The combination of AR and MA models results in a third class known asAutoRegressive Moving Average (ARMA) models, which are capable of representing weakly stationary time series, that is, series whose statistical properties such as mean and variance remain constant over time and are mathematically described as:

$$Y_t = c + \theta_1 Y_t + \theta_2 Y_{t-1} + ... + \theta_p Y_{t-p+1} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ... + \phi_q \varepsilon_{t-q}. \tag{2.12}$$

However, as some stochastic processes do not exhibit stationary characteristics, this limits the applicability of ARMA models. To address this limitation, the AutoRegressive

Integrated Moving Average (ARIMA) model was developed, with the inclusion of a differencing operation applied to the data, in order to convert non-stationary data into stationary data, thus making it possible to apply theARMA modelling framework. It is performed by applying a first-order difference to each data point:

$$Y_t' = Y_t - Y_{t-1} \tag{2.13}$$

This operation removes linear trends and effectively stabilizes the statistical properties of the time series, enabling the application of ARMA modelling techniques. The key advantage of transforming a dataset into a stationary form is the increased accuracy and reliability of future predictions, as stationary models tend to generalize better over time.

Therefore, the general notation for ARIMA models is given by ARIMA (p,d,q), where:

- p is the order of the AR part,
- d is the order of differencing required to achieve stationarity,
- q is the order of MA part.

The values for p and q can be determined with the help of the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF), which analyse the correlation structure of the time series, returning the influence of past values on a given observation. Specifically, the ACF is typically used to identify the appropriate order of the MA component, while the PACF is used to determine the order of the AR component.

The choice of differencing order d is influenced by the nature of the trend present in the data. A first-order difference is typically adequate for removing linear trends. In contrast, if the data display quadratic or cubic trends, second- or third-order differencing may be necessary to effectively remove these trends and obtain a stationary series.

To remove underlying patterns such as trends and seasonality, allowing for more accurate modeling and forecasting, detrenfing may be applied. Two common approaches to detrending are additive and multiplicative methods, which differ based on how the trend and seasonal components combine with the observed data. Additive detrending assumes that these components combine through simple addition, meaning the seasonal fluctuations have a constant magnitude regardless of the series level. This approach is appropriate when the variation in the data remains roughly constant over time, and the trend follows a linear or slowly changing pattern.

Conversely, multiplicative detrending assumes that the trend and seasonal components interact by multiplication. This means the size of seasonal fluctuations changes proportionally with the level of the series, often increasing as the series grows. Multiplicative models are suitable for time series exhibiting exponential growth or heteroscedasticity, where variance increases over time. Applying a logarithmic transformation is a common technique to convert multiplicative effects into an additive form, simplifying the modeling process.

2.4.4 Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors

The Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) model extends the ARIMA framework by incorporating both seasonal effects and external (exogenous) variables into the forecasting process, which makes it particularly well-suited for time series that exhibit seasonal trends and are influenced by other observed factors outside the series itself. In fact, a SARIMAX model can be viewed as an ARIMA model with additional seasonal terms and a regression component on external variables.

The general form of a SARIMAX model with exogenous regressors X_t can be expressed as:

$$\begin{aligned} y_t = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \Phi_1 y_{t-s} + \\ ... + \Phi_P y_{t-Ps} + \Theta_1 \varepsilon_{t-s} + ... + \Theta_Q \varepsilon_{t-Qs} + \beta^T X_t + \varepsilon_t \end{aligned} \tag{2.14}$$

Where ε_t denotes white noise, and β represents the coefficients associated with the exogenous regressors.

It is denoted by: $SARIMAX(p,d,q) \times (P,D,Q,s)$ where (p,d,q) are the non-seasonal orders corresponding to the autoregressive (AR), differencing (I), and moving average (MA) components, respectively and (P,D,Q,s) represent its seasonal counterparts, with s indicating the periodicity.

SARIMAX is particularly powerful in real-world settings where time series are influenced not only by their own past values but also by other observed factors. For example, in industrial wastewater treatment processes, effluent concentrations may depend on internal process dynamics (captured by ARIMA terms), seasonal operational patterns (captured by the seasonal component), and external inputs such as production rates or temperature (captured by exogenous variables).

The seasonal differencing operator is used to eliminate recurring patterns in time series data, enhancing stationarity and improving the model's suitability for forecasting. When exogenous variables are incorporated, they help capture external influences on the target series that univariate models might miss.

As with ARIMA models, the order selection for SARIMAX can be guided by the analysis of autocorrelation function plots, along with domain-specific knowledge of relevant seasonal cycles and potential external influences. The ACF plot displays bars at successive lags, where the height of each bar represents the correlation coefficient at that lag. When a bar extends beyond the significance bounds, it indicates significant autocorrelation at that specific lag. A gradual decline in bar heights typically suggests a long-term dependency within the time series data, while regular spikes occurring at fixed intervals point to the presence of seasonality. In contrast, PACF measures the correlation between observations at two time points after accounting for the influence of observations at all shorter lags, thereby isolating the direct relationship between those specific lags. The PACF plot is instrumental in identifying the order of an AR model. Significant spikes at particular lags in the PACF suggest which lags should be included in the AR model, and the point at which the partial autocorrelation cuts off helps determine the maximum lag order appropriate for the model specification.

Due to its ability to account for both seasonality and exogenous effects, SARIMAX is frequently more accurate than simple ARIMA models when such components are present in the data. Nonetheless, the inclusion of exogenous inputs requires careful attention to their quality, temporal alignment, and explanatory power.

2.4.5 Time Series Modelling in WWTP Systems

Time series modelling has become an indispensable tool for analyzing and forecasting the behavior of environmental and industrial processes, especially in the context of biological wastewater treatment. Variables such as dissolved oxygen, Chemical Oxygen Demand (COD), nutrient concentrations, and flow rates in WWTPs typically exhibit pronounced autocorrelation, seasonality, and process feedbacks, making classical statistical approaches inadequate for capturing the full spectrum of system dynamics (DÜRRENMATT; GUJER, 2012). Time series analysis, therefore, offers a more robust framework for modelling such dynamics, supporting both operational decision-making and regulatory compliance.

Among the suite of available time series methodologies, models from the autoregres-

sive family have found widespread application in industrial WWTP settings. since variables such as Dissolved Oxygen (DO), COD, nutrient concentrations, and flow rates in WWTPs typically exhibit pronounced autocorrelation, seasonality, and process feedbacks. ARIMA-based models, in particular, are valued for their ability to capture both stationary and non-stationary behaviors and to accommodate trend and seasonal patterns that frequently arise in treatment plant data. In the wastewater domain, Afan *et al.* (2024) have demonstrated how autoregressive and regression-based models outperform naive baselines in Biological Oxygen Demand (BOD) and COD prediction even in the presence of missing data, particularly when structured preprocessing and rolling-validation schemes are applied.

However, empirical studies consistently highlight that model reliability in the WWTP context is heavily contingent on rigorous data preprocessing and validation. This includes the meticulous handling of missing values, detection and correction of outliers, and alignment of disparate data streams, which is essential due to the prevalence of sensor faults, asynchronous sampling, and operational disruptions in industrial environments (AFAN *et al.*, 2024). Without addressing these data quality issues, model predictions can be heavily biased or misleading, thus undermining process observability and control. (NEWHART *et al.*, 2019)

Furthermore, time series model performance in environmental systems is not solely determined by fit to historical data. The importance of rigorous model diagnostics, such as residual analysis and the assessment of autocorrelation in model errors, to ensure that underlying assumptions, for instance, independence and normality of residuals, are satisfied (BOX *et al.*, 2015). These practices are particularly vital in safety-critical applications like WWTPs, where forecasting uncertainty is often expressed through prediction intervals and thus must be appropriately quantified and communicated for operational decision support (DÜRRENMATT; GUJER, 2012).

It is also recognized that, while ARIMA models perform robustly under quasi-steady-state and well-instrumented conditions, their effectiveness diminishes in the presence of process upsets, abrupt operational changes, or non-linearities typical of activated sludge processes. As such, several authors have explored the augmentation of classical ARIMA with exogenous variables (ARIMAX/SARIMAX models) (DÜRRENMATT; GUJER, 2012) or the fusion of time series models with data-driven machine learning techniques to better cope with the complexities of industrial WWTP systems (COSENZA *et al.*, 2025). These hybrid modelling strategies have been shown to improve short-term forecasting accuracy and resilience to unpredictable process

disturbances, contingent once again on the underlying data integrity.

To that end, time series modelling constitutes a powerful approach for short-term prediction, anomaly detection, and process monitoring in industrial wastewater treatment plants when implemented with robust diagnostics and within a comprehensive data governance framework. Nevertheless, the effectiveness of these models depends fundamentally on addressing data quality challenges and adopting systematic validation techniques specific to the operational realities of industrial WWTPs (AFAN *et al.*, 2024).

2.5 Analysed Parameters

DO is one of the most critical parameters in the biological treatment of wastewater. In activated sludge systems, DO directly supports the metabolism of aerobic microorganisms, which are responsible for oxidizing organic compounds and promoting nitrification. Therefore, the control of this variable is necessary to ensure efficient biological activity in aeration tanks, as low DO concentrations can lead to incomplete degradation of organic matter, inhibition of nitrifying bacteria, and the proliferation of undesired microorganisms. Conversely, over-aeration results in excessive energy consumption, often accounting for more than 50% of a treatment plant's total energy use (Metcalf & Eddy Inc. *et al.*, 2014).

Its concentrations in biological reactors are influenced by several factors, such as the oxygen transfer efficiency, which depends on aeration system design and mixing intensity; Temperature, since oxygen solubility in water decreases with increasing temperature; Organic and nitrogen loading rates, which determine microbial oxygen demand; Sludge retention time and biomass concentration. (QUAN *et al.*, 2012).

In many WWTPs, dissolved oxygen is monitored using online sensors and serves as a control variable for adjusting blower speeds or airflow rates. Given the non-linear dynamics of oxygen transfer and microbial activity, DO is commonly selected for predictive modelling using time series or machine learning approaches. Accurate DO forecasting enables the optimization of aeration, a primary energy consumer, and supports proactive operational control (Metcalf & Eddy Inc. *et al.*, 2014).

In this study, DO was selected as a target variable due to its operational importance, real-time availability via sensors, and its proven role as a key indicator of biological performance in the activated sludge process.

On the other hand, sulfate concentration is a key indicator of industrial load and

biological sulfate-reduction processes in wastewater treatment (XUE *et al.*, 2017). In anaerobic zones or poorly aerated systems, sulfate-reducing bacteria can convert sulfate into hydrogen sulfide, leading to odor issues, corrosion, and toxicity. In industrial WWTPs, sulfate loads often originate from external processes such as cleaning operations, chemical formulations, or high-sulfate influents (Metcalf & Eddy Inc. *et al.*, 2014). In conventional WWTP operations, particularly those treating industrial influents, sulfate is typically considered an exogenous variable due to its dependence on upstream discharges and its weak interaction with biological treatment processes (TCHOBANOGLOUS *et al.*, 2003).

Unlike dissolved oxygen, sulfate concentrations are not actively controlled in most WWTPs, and they tend to vary based on influent characteristics and slow, cumulative biological or chemical transformations. This makes sulfate particularly challenging to predict using purely endogenous plant operational data (BAHRAMIAN *et al.*, 2023). However, since it is one of the parameters monitored by the plant that currently are not meeting the required standards, it was chosen as a studied parameter in order to investigate the possibility of improving the performance of the WWTP solely by a data-driven approach.

2.6 Model Evaluation Metrics

In the context of statistical learning, model evaluation is a critical step to avoid both underfitting and overfitting, and to ensure that the selected model generalizes well to unseen data (HASTIE; TIBSHIRANI, 2009). Accordingly, in order to assess the predictive performance of the models developed in this study, the following evaluation metrics have been employed:

- Root Mean Square Error (RMSE): Defined as the square root of the average of the squared differences between the predicted and observed values, it penalizes larger errors more severely than smaller ones, being particularly sensitive to outliers and is useful in applications where large deviations from the true values are especially undesirable. Given a set of predictions $\hat{y}_i$ and true values y_i, it is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.15}$$

- Mean Absolute Average (MAE): Represents the average of the absolute differences between predicted and actual values. Unlike RMSE, MAE assigns equal weight to all errors, making it a more robust indicator when the data contains outliers or non-Gaussian noise.

It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.16}$$

- R^2-score (Coefficient of Determination): R^2 quantifies the proportion of variance in the dependent variable that is captured by the independent variables. It ranges from 0 to 1, where values closer to 1 indicate a better fit. In cases where the model performs worse than the mean of the target variable, R^2 may become negative. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.17}$$

  where $\bar{y}$ denotes the mean of the observed values.

Each of these metrics serves a specific purpose in evaluating model performance. While RMSE and MAE provide direct measures of prediction accuracy, R^2 offers a normalized assessment of model fit that is particularly relevant for linear regression frameworks.

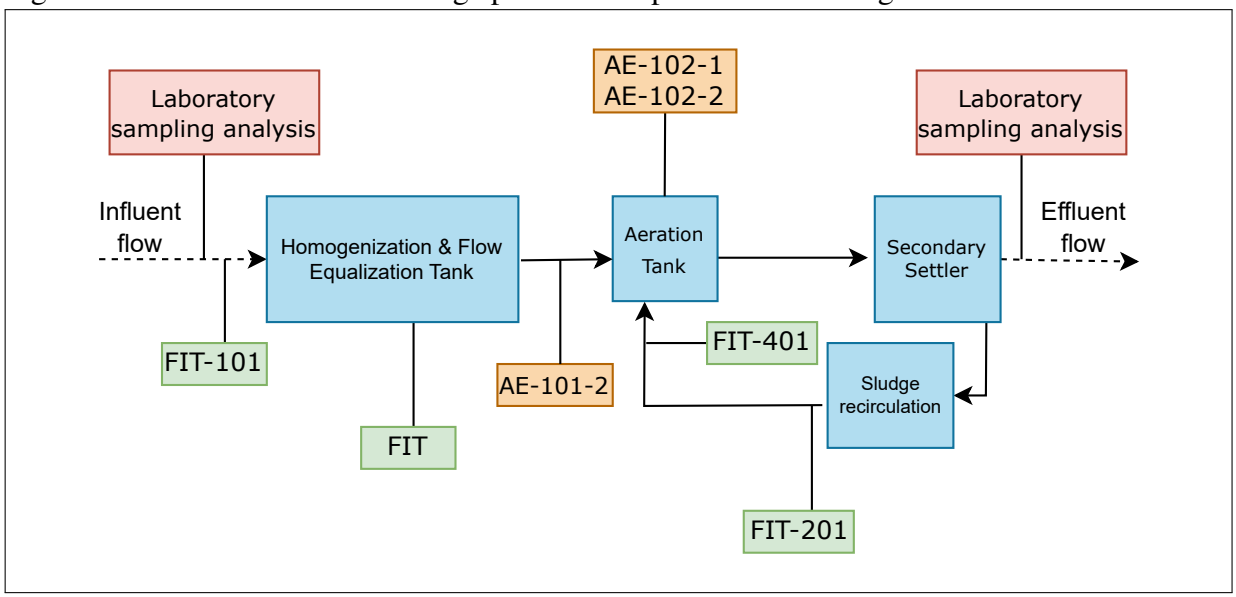
3 MATERIALS AND METHODS

3.1 Case Study Description

3.1.1 Overview of the Industrial WWTP

This study was conducted using operational data from an industrial-scale WWTP located in the state of Ceará, in Northeastern Brazil. The facility operates under an activated sludge configuration (Figure 3), processing a mixture of domestic and industrial effluents generated by textile manufacturing activities. The plant was selected for this study due to its complex operational environment, characterized by significant influent variability and partially automated monitoring infrastructure.

Figure 3 – WWTP's activated sludge process' simplified P and I diagram



Source: made by the author(2025).

The WWTP's process line comprises a raw wastewater pumping station, static screening, a homogenization and flow equalization tank (THRV) (Figure 4), primary flotation tank, an aeration tank (Figure 5), a secondary settler (Figure 6), sludge handling units, and final effluent discharge. The activated sludge process effectively occurs within the aeration tank and secondary settler, where it operates to remove organic matter, nitrogen, and other pollutants. Based on the operational priorities and data availability, this study focused exclusively on this region of the plant, following the approach recommended in previous works at the facility.

Figure 4 – Input flow into the homogenization and flow equalization tank

Source: made by the author (2025).

Figure 5 – Aeration tank during the process at the WWTP, filled with wastewater

Source: made by the author (2025).

Figure 6 – Secondary settler at the analysed WWTP

Source: made by the author(2025).

3.2 Data Acquisition

Data acquisition at the WWTP combines continuous online monitoring and periodic laboratory analyses. The dataset used in this study reflects both the heterogeneous and partially manual nature of data acquisition in industrial WWTPs, a factor carefully considered during model selection and evaluation. This research focused exclusively on variables associated with the aeration tank and secondary settler, the biological treatment core of the WWTP. (DÜRRENMATT; GUJER, 2012). The parameters and measurement points were identified using the facility's simplified Piping and Instrumentation Diagram (Figure 3) and operational records. Although the plant possesses several online sensors for process monitoring, namely, the flow indicator transmitters (FIT) (represented in green tags in the diagram and the analyser equipment (AE) for physico-chemical analysis (represented in orange tags) the system's operational routine relies largely on periodic laboratory analyses (represented in pink tags) and manual control adjustments and no integrated advanced process control is currently in place. For the influent laboratorial analysis, pH, Temperature, Centrifuge Residue, BOD, COD, and Nitrogen, Phosphate, Sulphate, Sulphide concentrations are sampled and analysed. As for the effluent analysis, pH, Sludge Bed Height, Centrifuge Residue, COD, BOD, Suspended Solids, color and turbidity and Total Nitrogen, Ammonium, Phosphate, Sulphate concentrations are evaluated. Although several influent and effluent parameters were available from laboratory monitoring records, the laboratory sampling was irregular. This inconsistency and sparseness of the data introduced substantial gaps, which would hinder robust time-series modelling if multiple variables were included. For this reason, only sulfate concentration was selected for inclusion in the analysis. This decision was also supported by the fact that sulfate levels frequently exceeded environmental regulatory limits, making it a particularly relevant and problematic parameter for investigation. The monitored parameters for this study are described in Table 1.

It is important to take into account that data for pH, temperature, and DO were extracted exclusively from the real-time analyzer records, as their equivalents in the laboratory dataset represent daily averages and lack sufficient temporal resolution for time series modeling. On the other hand, sulfate concentration data were obtained from laboratory records, characterized by a discrete daily sampling frequency, with some gaps present. No additional laboratory data were included in the analysis due to a high degree of inconsistency across datasets, as several parameters were measured only sporadically, often in months lacking any corroborating on-line data, which rendered them statistically unreliable. Consequently, incorporating such inputs

Table 1 – Monitored parameters in the WWTP analysed in this work

Variable	Tag	Unit	Source	Sampling Period
Influent flow rate	FIT-101	m^3/h	Online	10 minutes
Sludge recirculation flow rate	FIT-201	m^3/h	Online	10 minutes
Effluent flow rate	FIT-202	m^3/h	Online	10 minutes
Blower air flow rate	FT	m^3/h	Online	10 minutes
Dissolved oxygen (DO)	AE-102-2	mg/L	Online	1 minute
pH	AE-101-2	n.a.	Online	1 minute
Temperature (THRV)	AE-101-2	$^{\circ}C$	Online	1 minute
Temperature (Aeration Tank)	AE-102-1	$^{\circ}C$	Online	1 minute
Sulfate concentration	n.a.	mg/L	Laboratory	Daily

into the modeling process would be methodologically unsound and unlikely to give meaningful results.

3.3 Data Description and Preprocessing

A critical challenge in the analysis of operational data from industrial WWTPs lies in the integration of heterogeneous data sources, each with distinct sampling frequencies and reliability levels. In this study, actuator data were recorded every 10 minutes, while sensor data were collected at a higher frequency of one-minute intervals. Moreover, sulfate concentration, which should be obtained through daily laboratory analysis, exhibited significant sparsity, with only 17 values available over a 30-day period. This irregularity limits the ability to capture daily trends and introduces challenges in model training and validation, particularly when attempting to link slow-response variables such as sulfate to high-frequency operational signals.

3.3.1 Exploratory Data Analysis

Prior to model development, a comprehensive exploratory data analysis was conducted to understand the temporal dynamics, interdependencies, and statistical characteristics of the dataset in order to better to guide model selection and preprocessing strategies, by identifying stationarity issues, seasonality patterns, correlation structures, and potential causality between variables.

The analysis included:

- Time series plots of DO, pH, temperature and sulfate to visualize seasonal patterns, operational events, and anomalies.

- Boxplots and histograms to assess the distribution and identify outliers.
- Correlation analysis to examine relationships between variables.
- Identification of operational anomalies and sensor drift periods by cross-referencing analyzer data with operational notes and dosing system logs.

As estabilished by Orhon *et al.* (2009), the industrial WWTP environment is characterized by intermittent sensor faults, manual data entry errors, and asynchronous sampling routines, which result in missing data in both continuous and discrete records, issues that actually happened in this plant and affected the dataset. However, given the limitations inherent in the available dataset, in particular, the absence of systematic anomaly annotations or corrective procedures, it was necessary to implement a pragmatic strategy for ensuring data quality prior to model development.

Therefore, a careful analysis was conducted to identify time periods in which the process operated under relatively stable conditions, with minimal occurrence of anomalous values or operational interruptions. This involved visual inspection of time series plots as well as the application of basic descriptive statistics and anomaly detection techniques. Based on this assessment, the time interval of November 19th to December 17th, 2024 was selected for the analyzers' and actuators' datasets intersection, wherein the data exhibited consistent behavior with the laboratory data and the frequency of irregularities was notably reduced.

For sulfate sampling data, which presented a lower frequency, no imputation was performed, and, where required for model alignment, the Last Observation Carried Forward (LOCF) approach was applied to synchronize sulfate values with the continuous data. The LOCF is a commonly used technique in time series analysis when dealing with missing values. Its use in this case might be appropriate especially because it prevents the introduction of data where actually none exist, maintaining some continuity in the dataset. This selection process, although restrictive in terms of sample size, was essential to preserve the internal validity of the forecasting models and to prevent distortions caused by unfiltered anomalies.

3.3.2 Variable Selection and Data Filtering

In the context of multivariate time series modeling for industrial WWTPs, understanding the dynamic relationships among process variables is essential. These relationships inform the selection of predictors, help identify redundancies, and ensure that model assumptions, such as independence and stationarity, are met. All preprocessing steps were conducted

iteratively. Visual inspection and statistical confirmation were used at each stage to validate the results and ensure that the processed time series met the necessary assumptions for predictive modeling.

- **Correlation Matrix:** To explore feature correlation and causality, a correlation matrix was visualized to assess linear dependencies between variables. The Pearson correlation coefficient was used to construct the matrix, as it provides a standardized measure of linear dependence between two continuous variables. The coefficient ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with values near zero indicating weak or no linear relationship.
- **Augmented Dickey-Fuller (ADF) test** : Prior to the application of Granger causality tests and the development of time series models, all candidate variables were subjected to a stationarity assessment using the Augmented Dickey-Fuller (ADF) test. The test was applied individually to each time series variable in the dataset spanning the period from October to December. These variables included influent flow rates, COD concentrations, total suspended solids, and other operational metrics of the wastewater treatment process. Many operational variables, including dissolved oxygen and temperature, were either already stationary or made stationary through first-order differencing. To complement the statistical tests, rolling statistics (mean and standard deviation over time) were plotted to visually examine trends and potential seasonality.
- **Additive and Multiplicative Decompositions and Seasonal Detrending:** Aditionally, given the cyclical nature of biological wastewater treatment processes and the evidence of non-stationarity in several variables, additive and multiplicative decompositions were performed to isolate trend, seasonal, and residual components, in order to extract trend and seasonal effects that may confound modeling efforts, and to isolate the residual or stochastic component for use in predictive models, given that the variables are often governed by overlapping dynamics. To address these aspects, we applied classical decomposition techniques, treating each variable as the sum (additive model) or product (multiplicative model) of these components.After that, detrending was implemented by subtracting the line of best fit and removing the trend component from decomposition results.
- **Granger Causality Test:** In order to understand dependencies among variables, select predictors for multivariate time series models and assess directionality in relationships between series, Granger causality tests were employed to evaluate whether lagged values

of one variable could significantly improve the prediction of another, thereby informing the specification of models such as VAR and SARIMAX. Following stationarization, Granger tests were conducted between key process variables and the targets (e.g., flow vs. oxygen) to infer potential predictive relationships and causality directions. Lag values from 1 to 5 were tested. The selection of lag order was based on domain knowledge and supported by the Bayesian Information Criterion (BIC) where applicable. For each pair of variables (X_t, Y_t), the null hypothesis H_0, that X_t does not Granger-cause Y_t, was tested using an F-test comparing restricted and unrestricted vector autoregressive models.

3.4 Modelling

The objective of this study was to develop data-driven forecasting models for key performance indicators in the WWTP's activated sludge process, specifically targeting Dissolved Oxygen (DO) and sulfate concentrations. Based on data availability and process dynamics, four modeling techniques were selected: VAR, ARIMA, SARIMAX, and Multiple Linear Regression (MLR).

3.4.1 Model Selection

- **Multiple Linear Regression (MLR):** Applied for DO and sulfate prediction using operational and environmental variables as predictors. Although linear, MLR offers transparency and interpretability, characteristics valued in operational environments with limited analytical infrastructure (DÜRRENMATT; GUJER, 2012).
- **VAR:** A VAR model was applied to the sulfate variable because of its conceptual similarity to standard multiple regression models and the relative simplicity involved in fitting them to empirical time series data as model parameters can be estimated using least squares methods, which yield closed-form solutions and are computationally efficient.
- **ARIMA:**A widely applied statistical model for univariate time series forecasting, capable of capturing trends and autocorrelation structures in continuous process data, ARIMA is suitable for variables such as DO, whose temporal behavior is influenced by internal process feedbacks and operational adjustments (AFAN *et al.*, 2024).
- **SARIMAX:** An extension of ARIMA that incorporates external explanatory variables, enabling the inclusion of variables such as temperature and pH when predicting DO

or sulfate concentrations, SARIMAX accommodates both autoregressive and seasonal patterns and is recommended for WWTP processes influenced by multiple interacting factors (BAHRAMIAN *et al.*, 2023), thus it was also applied to DO forecasting.

3.4.2 Evaluation and Tuning Strategies

According to the principles of statistical learning theory, models were evaluated on independent test data whenever possible to ensure the reliability of performance estimates and to assess the generalization error beyond the training sample (HASTIE; TIBSHIRANI, 2009). The simultaneous use of multiple metrics (RMSE, MAE, R^2) allows for a more nuanced understanding of model strengths and limitations under different operational scenarios.

Aditionally, in accordance with best practices for time series modeling (NEWHART *et al.*, 2019), a rolling-origin cross-validation scheme was implemented. This method respects the temporal structure of the data and prevents information leakage from future observations into past model estimates.

The procedure consisted of:

1. Defining an initial training window.
2. Fitting the model on the training set.
3. Predicting the next observation (forecast horizon = 1 step).
4. Rolling the training window forward by one observation and repeating the process.

For each iteration, performance metrics were computed, and aggregated results were then used to assess model robustness and predictive accuracy.

3.5 Computational Environment

All analyses were implemented on a local Windows machine using a Python 3.13 virtual environment, leveraging the pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels, and scipy libraries. The computational notebook was developed using Jupyter and maintained with clear version control and commenting for reproducibility.

4 RESULTS

This chapter discusses the comparative performance of these models, their interpretability, and the broader implications for WWTP monitoring and optimization. The results of this study reveal the strengths and limitations of data-driven modeling in the context of an industrial wastewater treatment plant that operates under real-world constraints such as sensor limitations, influent variability, and imperfect data quality. The analysis focused on two key performance indicators: dissolved oxygen and sulfate concentration, each modeled using distinct statistical approaches.

4.1 Exploratory Data Analysis

Throughout the preprocessing phase, several opportunities for data quality enhancement were identified, including missing or null readings during sensor calibration or operational downtime, along with gaps present in the laboratory records due to inconsistencies in recording frequency, which did not always adhere to the recommended operational standards. These variations, while common in industrial settings, present valuable challenges for further refinement and affected model calibration and forecasting accuracy, particularly in periods with incomplete or inconsistent data. While the available data provided a valuable basis for modeling, the experience highlights the importance of structured data management protocols, systematic operator training, and gradual investment in integrated automation, as recommended by Bahramian *et al.* (2023).

4.1.1 Initial dataset overview

The raw data revealed several inconsistencies, such as falling outside the physical range of the variables. (Table 2). For instance, in pH measurement, an anomalous minimum of -17 was observed, which suggests sensor malfunction or data recording errors. Likewise, a set of negative values were observed for DO data (Figure 7), further suggesting equipment error.

Similarly, temperature reading exhibited large variations, including unfeasible values such as negative temperatures, even though the mean and median values were in consonance with an environment exposed to a warmer climate Additionally, for a considerable amount of time, the values were read as constant, an overall behaviour which further suggests potential data errors and sensor malfunctions. 8

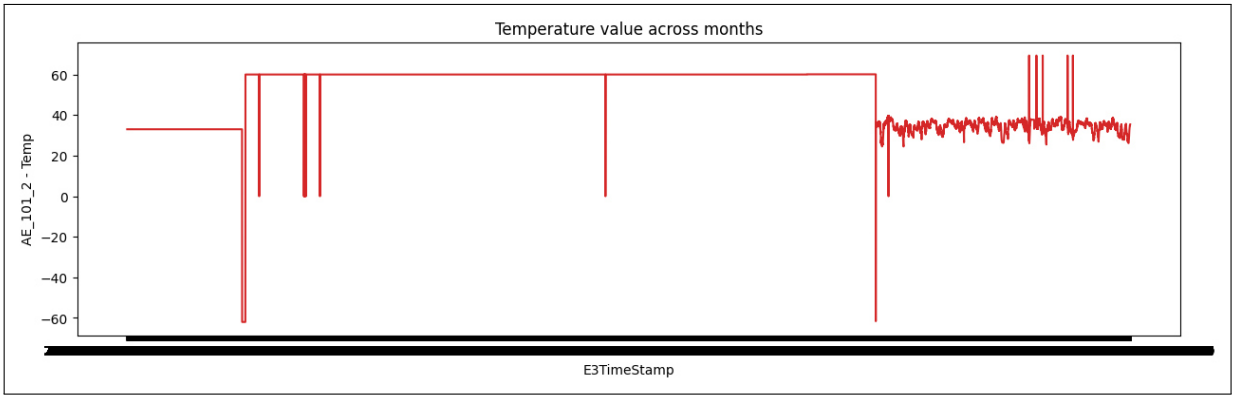
Table 2 – Descriptive statistics for AE raw sensor data

	AE-101-2 - pH	AE-101-2 - Temp	AE-102-2 - Oxi	AE-102-1 - Temp
Mean	8.917	49.989	3.568	37.794
Standard deviation	1.644	14.229	2.457	2.372
Minimum value	-17.000	-62.000	-142.000	-71.000
Q1	8.896	34.935	2.816	36.745
Median	9.000	60.000	4.000	37.969
Q3	9.328	60.000	4.994	39.000
Maximum value	16.182	69.355	45.729	50.341

Figure 7 – DO box plot from raw data

Source: made by the author (2025).

Figure 8 – Temperature raw data along time



Source: made by the author (2025).

Table 3 – Descriptive statistics forfiltered WWTP sensor data (November to December)

	pH (THRV)	Temp (THRV)	Diss. O_2	Temp (Aer. Tank)
Mean	9.144	34.348	3.334	37.532
Standard deviation	0.496	2.661	1.012	1.025
Minimum value	7.848	25.687	0.006	34.870
Q1	8.810	32.908	2.806	36.877
Median	9.209	34.918	3.531	37.398
Q3	9.462	36.321	4.021	38.153
Maximum value	10.095	38.954	5.223	40.252

Table 4 – Descriptive statistics for flow rates and sulfate concentration (November to December)

	Entry Flow	Sludge Recirc.	Blower Flow	Sulfate
Mean	36.864	51.184	7172.179	1118.915
Standard deviation	21.143	6.358	1054.216	205.209
Minimum value	0.000	0.000	0.000	706.000
Q1	20.900	50.000	6727.250	1044.000
Median	35.200	50.000	7247.000	1181.000
Q3	49.825	50.100	7430.250	1296.000
Maximum value	185.600	86.000	9267.000	1340.000

4.1.2 Post-Data Cleaning Analysis

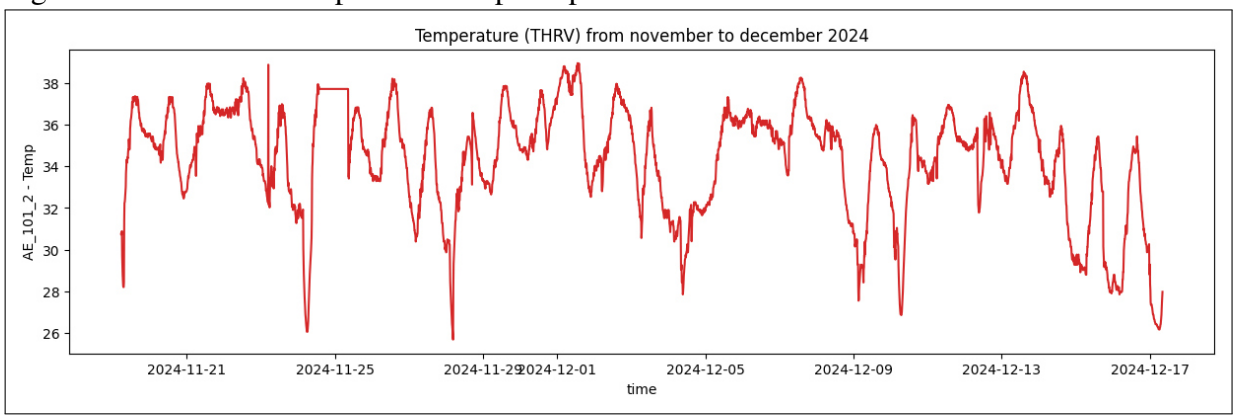
To address these isssues, time periods with evident sensor failures, constant readings, or excessive missing data were excluded from modeling. Aditionally, data acquisition frequency was reduced to one row every 10 minutes in order to make it compatible with the actuator's data. After this data cleaning process (Table 3), we are left with 4048 observations across four features related to sensor readings. The fact that the same rows were eliminated during the process where only physical chemical rules were estabilished to clean the data suggests there was a general system fault at the point when the outliers ocurred.

For the pH, distribution is approximately normal and tightly centered around 9.2, with a range is from 7.8 to 110 and a low standard deviation (0.496), suggesting typical pH values for basic environments, which are consistent observations for sludge samples.

For the temperature analysis, the THRV values' span indicaties a mix of cold and hot samples, likely due to environmental or operational changes, probably due to industrial influx at times. On the aeration tank, we see a symmetric distribution with a median close to the mean, around $37.81°C$, indicating a stable process, which is expected since it is open to the environment, and the wheather in the region at this time of the year is quite stable.

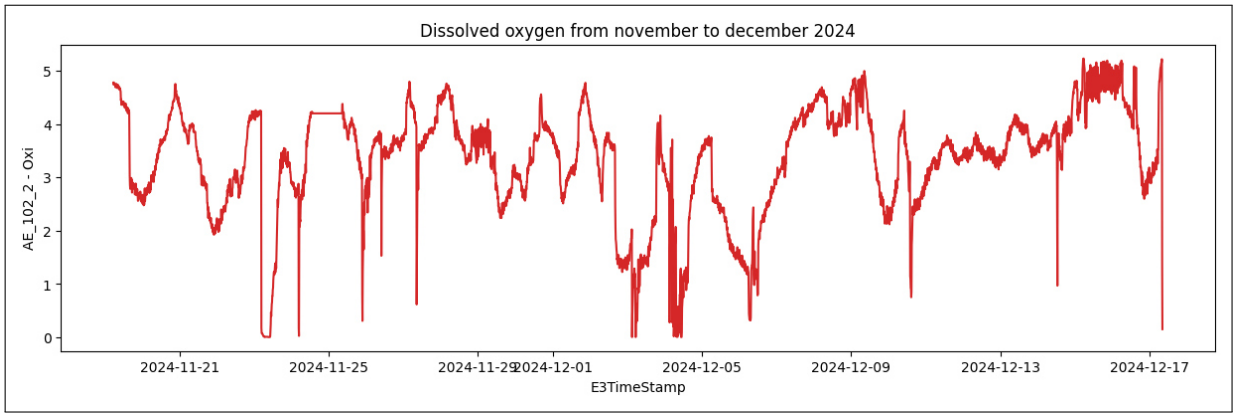
The DO time series (Figure 10) shows levels fluctuating mostly between 2 and 5 mg/L, which is typical for aerated biological treatment processes. This range suggests that for

Figure 9 – AE-101-2 temperature temporal plot



Source: made by the author (2025).

Figure 10 – Dissolved oxygen's temporal plot



Source: made by the author (2025).

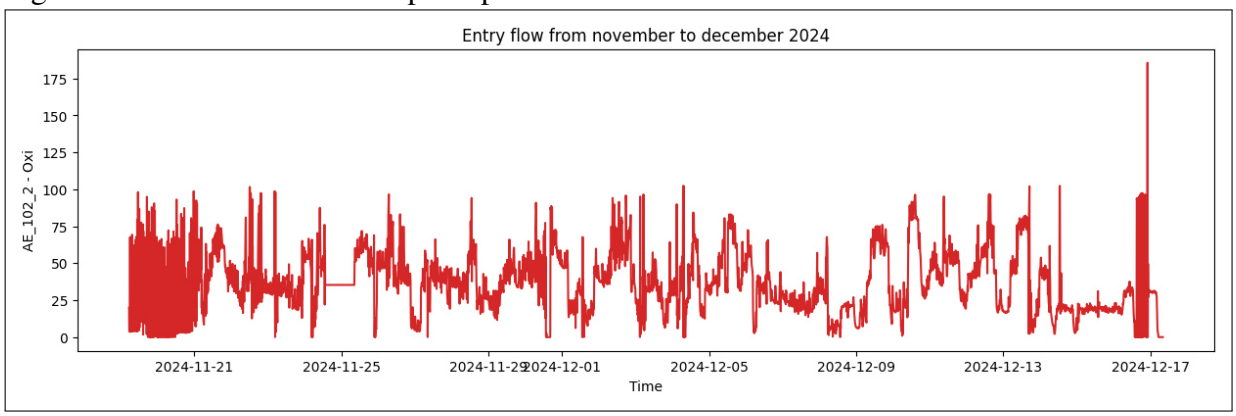
much of the period, aeration was active and responsive, though punctuated by abrupt variations corresponding to maintenance activities and operational interventions, as cross-checked against operator notes.

Regarding the influent's flow data 11, there are several intervals, particularly around November 21th, December 6th, and mid December where the flow drops to near-zero, which could correspond to a myriad of reasons, from industrial production shutdown to operational intervention. Nevertheless, its irregularity presents a major challenge for forecasting, especially for variables like sulfate concentration that depend heavily on incoming load.

The blower flow rate data (Figure **??**) reveals a predominantly stable operation, with average values ranging between 7000 and 8000 units, indicating consistent aeration system performance. However, several abrupt drops to near-zero values are observed, particularly around November 23–24, November 27, and December 3–4. These events likely correspond to transient instrumentation failures, brief blower shutdowns, or scheduled maintenance activities.

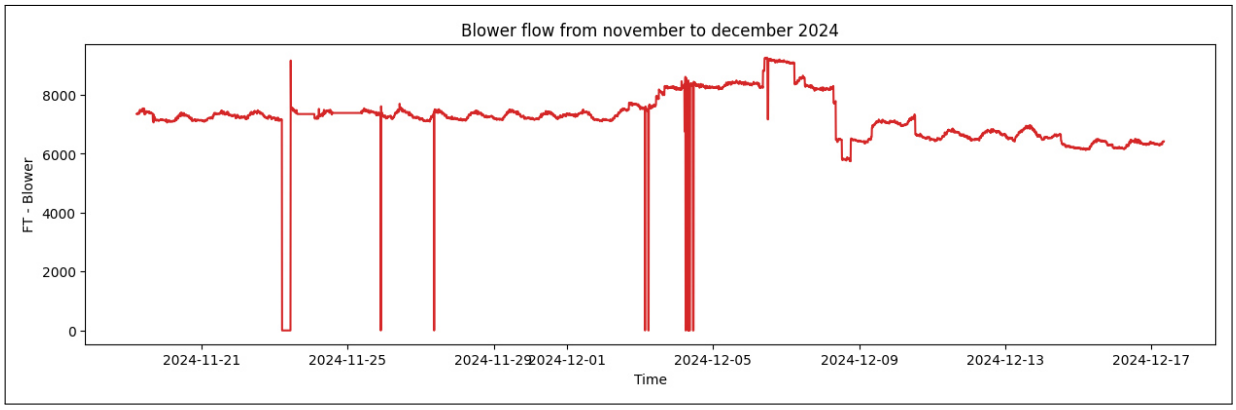
To ensure alignment between the laboratory and online monitoring data, the period

Figure 11 – Influent's flow temporal plot



Source: made by the author (2025).

Figure 12 – Blower's flow temporal plot



Source: made by the author (2025).

chosen for sulfate analysis corresponded to the most reliable operational phase of the online sensors. However, only 17 laboratory sulfate samples were available during this timeframe (Figure 13), and these displayed high variance, reflecting the irregular nature of industrial discharges, showing episodic spikes.

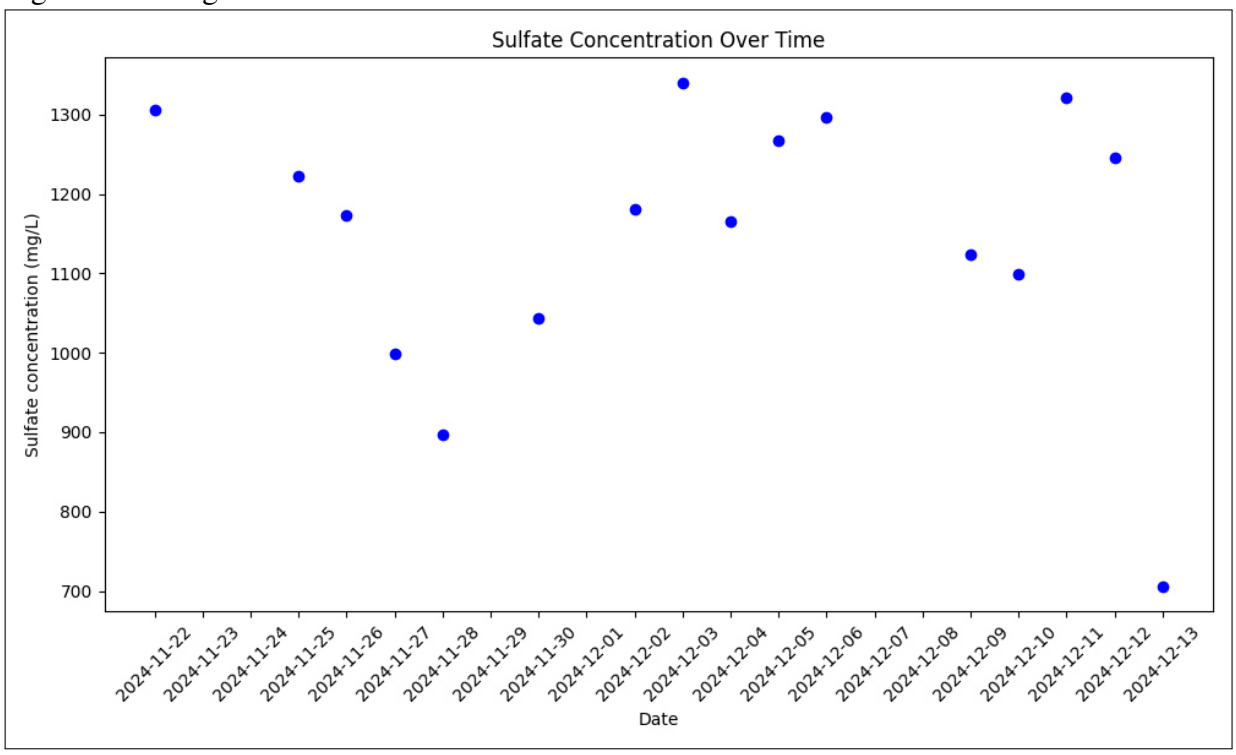
To address the resulting data gaps and enable consistent modelling, the LOCF method was applied for imputation (Figure 14).

4.1.3 Preprocessing Results

4.1.3.1 Correlation Matrix and Feature Selection

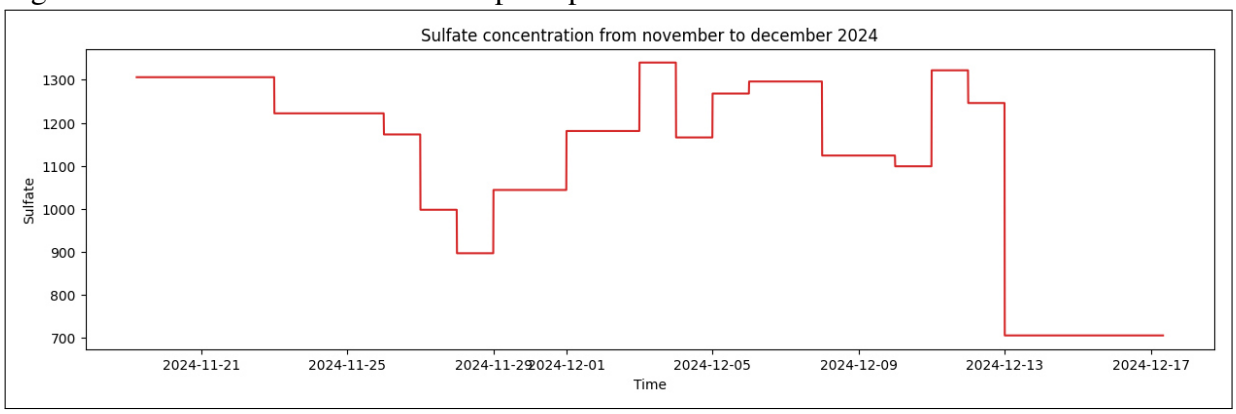
To enhance interpretability, the correlation matrix was visualized using a heatmap (Figure 15), which provided a clear insight into the dynamic relationships within the wastewater treatment process, particularly regarding sulfate and dissolved oxygen concentrations. In the context of this study, strong positive correlations were observed, for example, between influent DO and flow rate. These relationships are consistent with the physical and operational character-

Figure 13 – Registered data for sulfate concentration



Source: made by the author (2025).

Figure 14 – Sulfate concentration temporal plot after LOCF

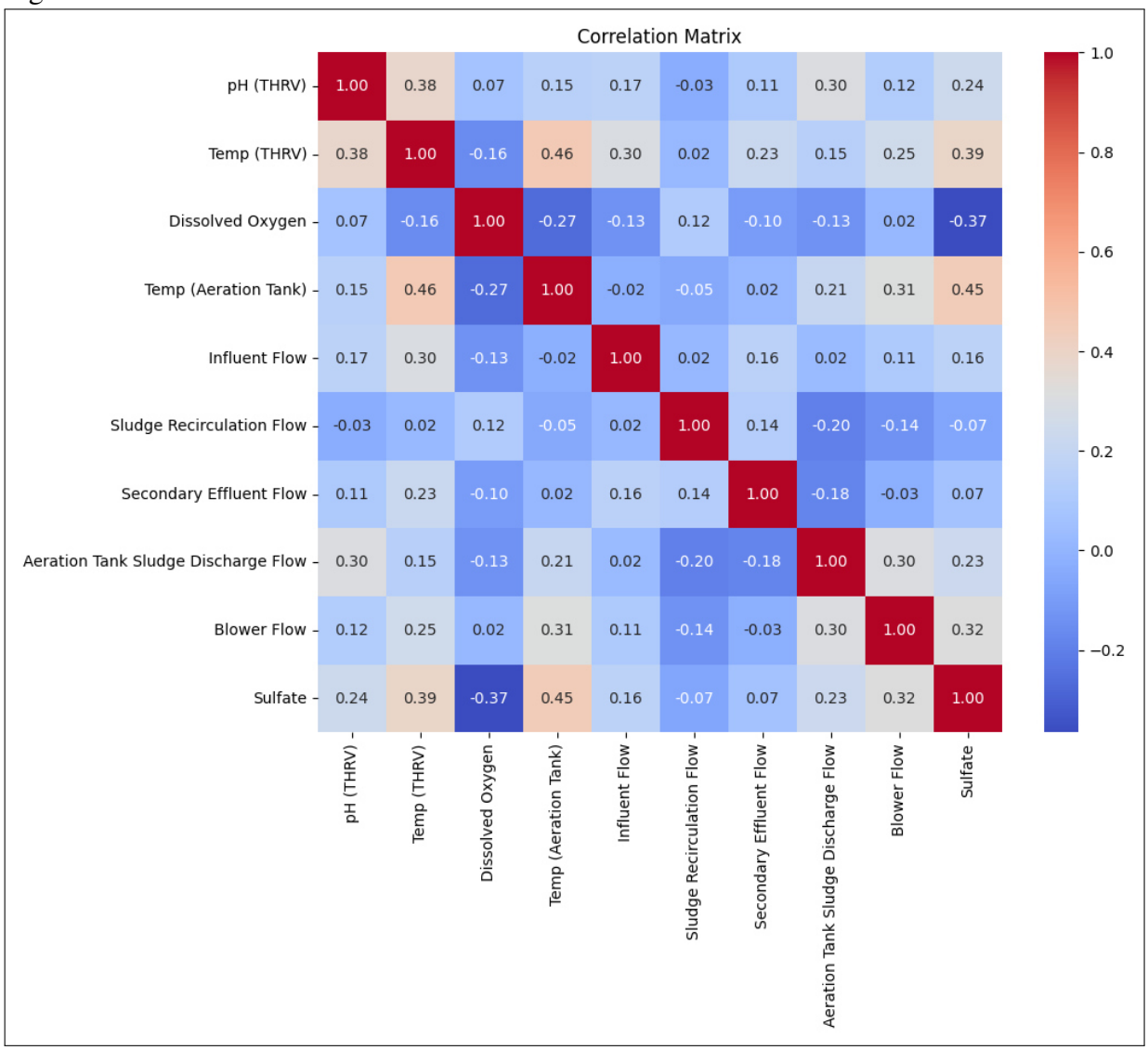


Source: made by the author (2025).

istics of the wastewater treatment process and reinforce the use of these variables in subsequent predictive modeling stages.

The observed moderate positive correlation between sulfate and temperature in the aeration tank suggests that sulfate accumulation may be thermally influenced, potentially due to increased microbial or chemical reaction rates under warmer conditions. Similarly, the positive correlation with blower flow may reflect the role of aeration in chemical conditions that affect sulfate transformations. Interestingly, sulfate was negatively correlated with dissolved oxygen, implying that oxygen-limited environments might favor sulfate formation or inhibit its removal. These relationships are consistent with the biochemical pathways of sulfur cycling and reinforce

Figure 15 – Correlation Matrix for the observed dataset



Source: made by the author (2025).

the inclusion of temperature, aeration, and oxygen availability as key drivers in sulfate modeling.

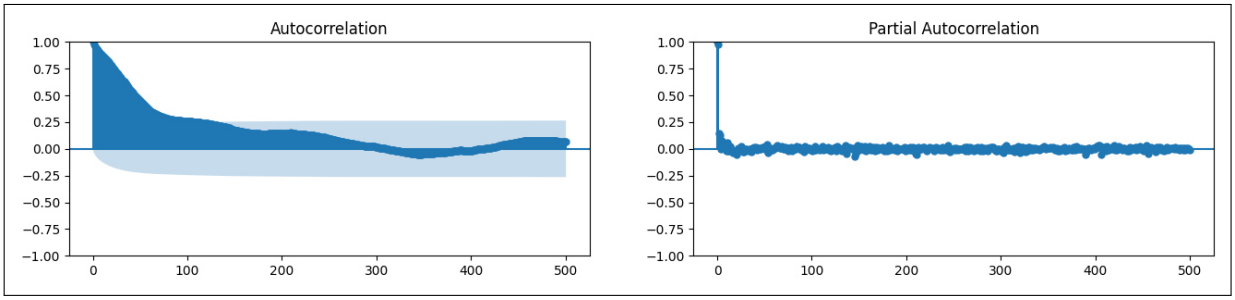
In contrast, the factors influencing dissolved oxygen appear to be less directly correlated with the broader set of variables. The inverse relationship with temperature aligns with known physical properties, specifically, the decreased solubility of oxygen at higher temperatures. Weak correlations with influent flow and sludge discharge suggest possible dilution or hydraulic effects, but their explanatory power remains limited. Notably, several variables originally considered, such as blower flow, sludge recirculation, and effluent flow showed minimal association with dissolved oxygen and were excluded from further analysis in order to reduce dimensionality and prevent overfitting due to the inclusion of noise.

4.1.3.2 Autocorrelation

According to the autocorrelation and partial function plots (Figure 16) DO shows characteristics consistent with an autoregressive process, as the series shows a gradual exponential decay in autocorrelation, suggesting persistent influence from past values. The PACF reinforces this observation, displaying a significant spike at lag 1 followed by an immediate drop-off. This pattern is indicative of an AR(1) process, in which the current value is primarily influenced by its immediate past. Such behavior suggests a level of short-term memory in the DO data, where recent observations exert a strong effect on future values, underscoring the need for autoregressive modeling to accurately capture the temporal structure.

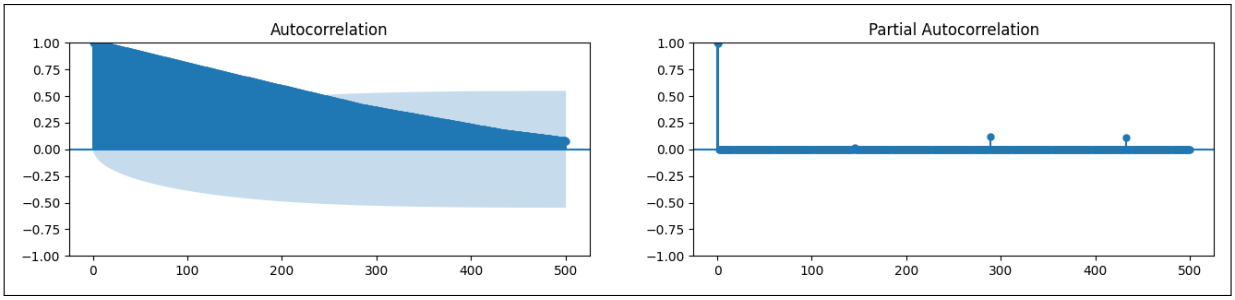
In contrast, the sulfate time series presents a more pronounced trend and persistence over time, as evidenced by its ACF plot (Figure 17), which declines very slowly and nearly linearly, which is expected given its sparser time series format. This long memory structure points to potential non-stationarity in the data. The PACF shows a sharp spike at lag 1, with all subsequent lags near zero, suggesting an underlying AR(1) process within a non-stationary framework. The observed autocorrelation structure indicates that sulfate concentrations are heavily influenced by previous values and possibly external factors contributing to a trend over time.

Figure 16 – ACF and PACF plots for dissolved oxygen



Source: made by the author (2025).

Figure 17 – ACF and PACF plots for sulfate concentration



Source: made by the author (2025).

4.1.3.3 Stationarity Testing Results

The Augmented Dickey-Fuller (ADF) test was applied to assess the stationarity of the time series data.

Table 5 – Results of the Augmented Dickey-Fuller Test for different variables

	pH	**THRV T.**	**Aeration T.T.**	**Influent**	**Blower**	**Sulfate**	**DO**
Test Statistic	-2.201	-5.930	-2.744	-7.769	-7.237	-1.331	-5.080
p-value	0.206	2.39e-07	0.067	9.03e-12	1.93e-10	0.615	1.5e-05
Lags Used	10	14	17	25	31	0	19
Observations	4037	4033	4030	4022	4016	4047	4028
Critical Value (1%)	-3.432	-3.432	-3.432	-3.432	-3.432	-3.432	-3.432
Critical Value (5%)	-2.862	-2.862	-2.862	-2.862	-2.862	-2.862	-2.862
Critical Value (10%)	-2.567	-2.567	-2.567	-2.567	-2.567	-2.567	-2.567

The test results indicate that the variables THRV temperature, Influent flow, Blower flow, and DO have test statistics well below the critical values and very low p-values, leading to rejection of the null hypothesis of non-stationarity. This result justifies the use of modeling techniques that assume stationarity in these analysed data series. Conversely, pH, Aeration tank temperature, and Sulfate concentration show test statistics that are not sufficiently low and p-values above common significance levels, indicating failure to reject the null hypothesis and suggesting non-stationarity for these series.

The resulting stationary series were then used as input for Granger causality analysis and the construction of predictive models.

4.1.3.4 Granger Test Results

Given the results of the Augmented Dickey-Fuller tests, it was observed that some variables in the dataset are stationary while others are non-stationary. Therefore, differentiation was applied when necessary.

Table 6 – Granger Causality Test Results for Influent Flow and Dissolved Oxygen

Lags	**SSR-based F**	**p-value**
1	12.2465	0.0005
2	11.1789	0.0000
3	8.2965	0.0000
4	6.7590	0.0000
5	6.0500	0.0000

Results for the Granger Test between Influent Flow and Dissolved Oxygen (Table 6)

indicate that for all lag lengths tested, the null hypothesis is strongly rejected, with test statistics indicating robust evidence of Granger causality, demonstrating that the explanatory variable contains significant information useful for predicting the dependent variable, up to five time periods in the past.

Table 7 – Granger Causality Test Results for Blower Flow and Dissolved Oxygen

Lags	SSR-based F	p-value
1	1.0892	0.2967
2	2.4778	0.0841
3	3.1561	0.0238
4	3.5704	0.0065
5	2.1822	0.0534

As for the Granger Test between Blower Flow and Dissolved Oxygen, the null hypothesis in each case is that past values of blower flow do not Granger cause dissolved oxygen. For lag 1, the F-statistic is 1.0892 with a p-value of 0.2967, indicating no statistical evidence of causality at conventional significance levels. Lag 2 yields a lower p-value (0.0841), but still fails to meet the standard 5% threshold. Starting from lag 3, however, the p-value drops to 0.0238, providing sufficient evidence to reject the null hypothesis at the 5% level. Lag 4 further strengthens this result, with a p-value of 0.0065, indicating a more robust causal relationship. At lag 5, the p-value increases to 0.0534, just above the 5% threshold, weakening the evidence again.

These results suggest that blower flow Granger causes dissolved oxygen when lags of 3 or 4 are considered, but not at shorter or longer lag lengths. Therefore, including 3 or 4 lags of blower flow in a predictive model of dissolved oxygen may provide useful information, while fewer or more lags may fail to capture the underlying dynamics.

Table 8 – Granger Causality Test Results for Influent Flow and Sulfate Concentration

Lags	SSR-based F test	p-value
1	0.0354	0.8507
2	0.0307	0.9698
3	0.0259	0.9944
4	0.0437	0.9964
5	0.0388	0.9992

However, for the sulfate Granger tests with the influent flow variable (Table 8), chosen for displaying the highest correlation with it (Figure 15), the F-statistics are extremely low for all lag lengths, and the associated p-values are substantially above significance levels,

indicating a complete lack of predictive power from influent flow to sulfate concentration. These results strongly suggest that autoregressive models with the sulfate time series will perform poorly.

4.2 Model Performance Results

The predictive performance of implemented models (VAR, ARIMA, SARIMAX, and Multiple Linear Regression) was evaluated using a rolling-origin cross-validation strategy.

4.2.1 Dissolved oxygen prediction

Dissolved oxygen was the most successfully modeled variable in this study.

The ARIMA(1,1,1) model (Table 9) estimated for the differenced series of dissolved oxygen yielded a strong in-sample fit, with a notably low Akaike Information Criteria (-580.39). Both the autoregressive ($AR(1) = 0.48$, $p < 0.001$) and moving average ($MA(1) = 0.64$, $p < 0.001$) components were highly significant, indicating that short-term dynamics in the series are well-captured by the model's structure. However, the residuals exhibited heavy left skew, as well as significant heteroskedasticity, indicating that the residuals of the ARIMA(1,1,1) model exhibit time-varying volatility. This suggests that while the model adequately captures the mean behavior of dissolved oxygen dynamics, it fails to fully account for periods of increased or decreased fluctuation in the data.

Table 9 – Coefficients from ARIMA(1,1,1) model for the seasonal component

Parameter	**Coefficient**	**Std. Error**	***p*-value**
AR(1)	0.4775	0.0290	<0.001
MA(1)	-0.6425	0.0270	<0.001
Variance (σ^2)	0.0477	0.0002	<0.001

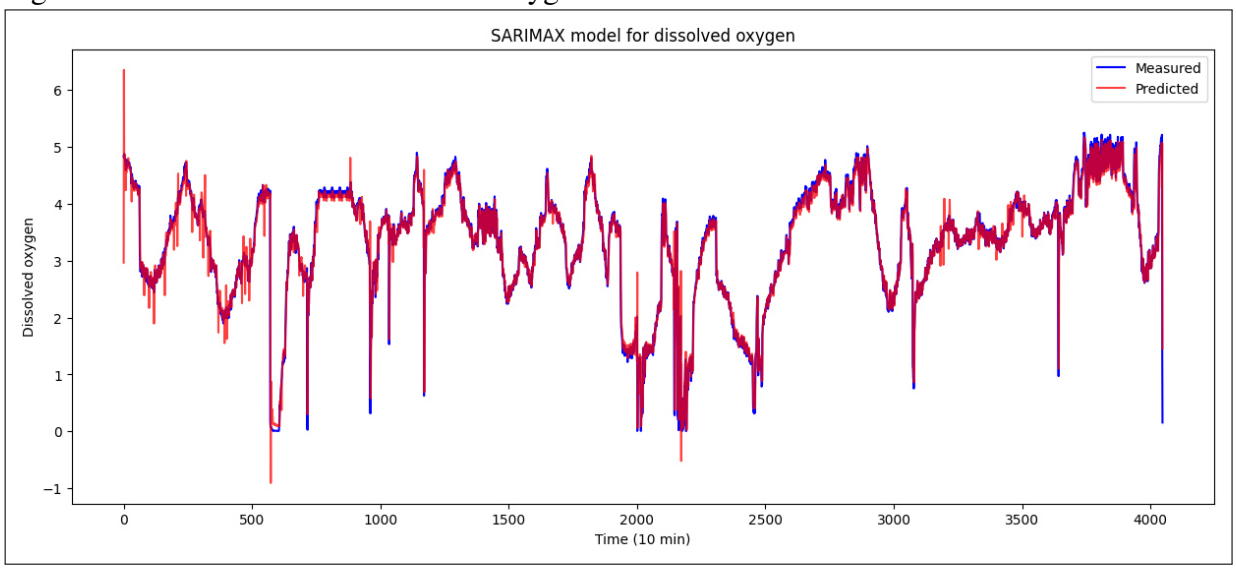
A SARIMAX(1,1,1) model (Figure 18) was then estimated using only the two most relevant exogenous variables: pH and influent flow. This decision was informed by results from an initial fuller model, in which the remaining exogenous variables exhibited high p-values ($p > 0.4$), suggesting weak or spurious contributions to the model. By retaining only the components with theoretical and statistical support, the reduced model avoids the risk of overfitting while maintaining explanatory clarity. In particular, influent flow remained a highly significant predictor ($p < 0.001$), highlighting its key role in shaping oxygen availability through both dilution effects and loading impacts.

Aditionally, linear regression models without the autoregressive component were tested. To ensure optimal performance and avoid overfitting, a pipeline was implemented to systematically evaluate a range of λ for both Ridge and Lasso regression models. The goal was to identify the best-fitting model by selecting the regularization parameter that minimized validation error while maintaining generalizability. Multiple candidate values for alpha were tested using cross-validation, and the model with the best overall performance metrics was retained. This approach allowed the selection of the most appropriate balance between bias and variance, enabling each model to prioritize meaningful predictors while penalizing noise and collinearity. The final Ridge model, using the best-performing parameter set, was then evaluated against the measured dissolved oxygen data and used for interpretation and comparison with time series-based alternatives. The Ridge regression model (Figure 19) for dissolved oxygen yielded a relatively smooth prediction profile, capturing the central trend of the time series despite failing to replicate its rapid fluctuations. With an R^2 score of only 0.122, the model indicated that linear relationships among the selected predictors were insufficient to account for the complexity of the dissolved oxygen dynamics. Moreover, its inability to track sharp rises and drops highlights the limitations of static linear approaches in modeling highly variable and autocorrelated environmental processes. Therefore, among the evaluated approaches (Table 10), the SARIMAX model achieved the lowest mean absolute error, indicating strong short-term predictive accuracy, despite a relatively higher root mean square error, which may reflect occasional larger deviations. The ARIMA model displayed a comparable performance, with a slightly higher MAE but lower RMSE, suggesting a more consistent error distribution. In contrast, Ridge Regression, even after hyperparameter tuning, exhibited considerably worse performance. These results reinforce the notion that autoregressive models are more suited to capturing the temporal structure and dependencies in dissolved oxygen dynamics, while linear regression struggles to explain its variability when relying solely on instantaneous process variables as predictors, although Ridge regularization helped mitigate overfitting by constraining coefficient magnitude.

Table 10 – Model performance for dissolved oxygen prediction

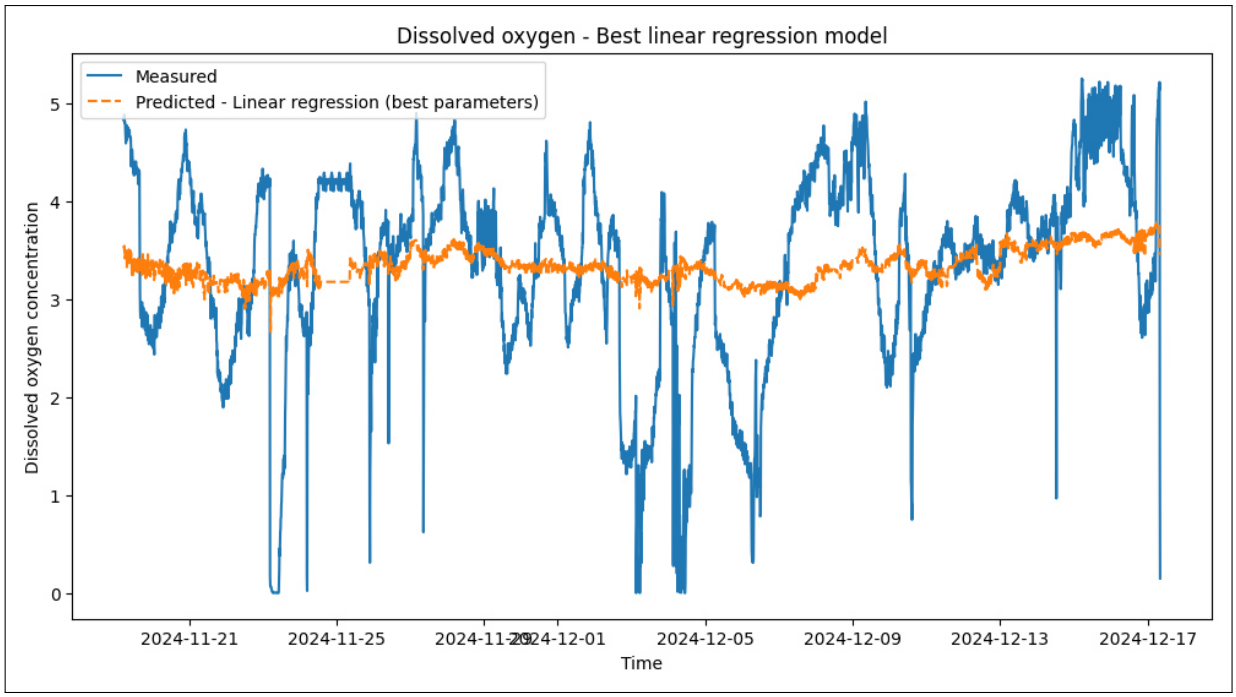
Model	**MAE**	**RMSE**	R^2
ARIMA (1,1,1)	0.196	0.183	–
SARIMAX (1,1,1)	0.110	0.231	–
Ridge Regression (λ=10000)	0.722	0.937	0.122

Figure 18 – SARIMAX for dissolved oxygen



Source: made by the author (2025).

Figure 19 – Ridge regression for dissolved oxygen



Source: made by the author (2025).

When comparing the AR models to the linear regression models used in this study, an important trade-off between model complexity and explanatory power emerges. AR models, tend to incorporate temporal dependencies by including lagged terms, which can improve fit on the training data. However, increasing the number of autoregressive terms risks overfitting, where the model captures noise rather than underlying patterns, potentially reducing generalization to unseen data. On the other hand, linear regression models (Figure 19) exhibited low R^2 values, suggesting limited ability to explain variability in the dependent variable based on the

chosen predictors, which may reflect the simplicity of these models or the omission of important temporal dynamics that AR models attempt to capture. Thus, while AR models risk overfitting through complexity, linear regressions might underfit by ignoring temporal structure, highlighting the need for a balanced approach that accounts for both time dependencies and parsimony.

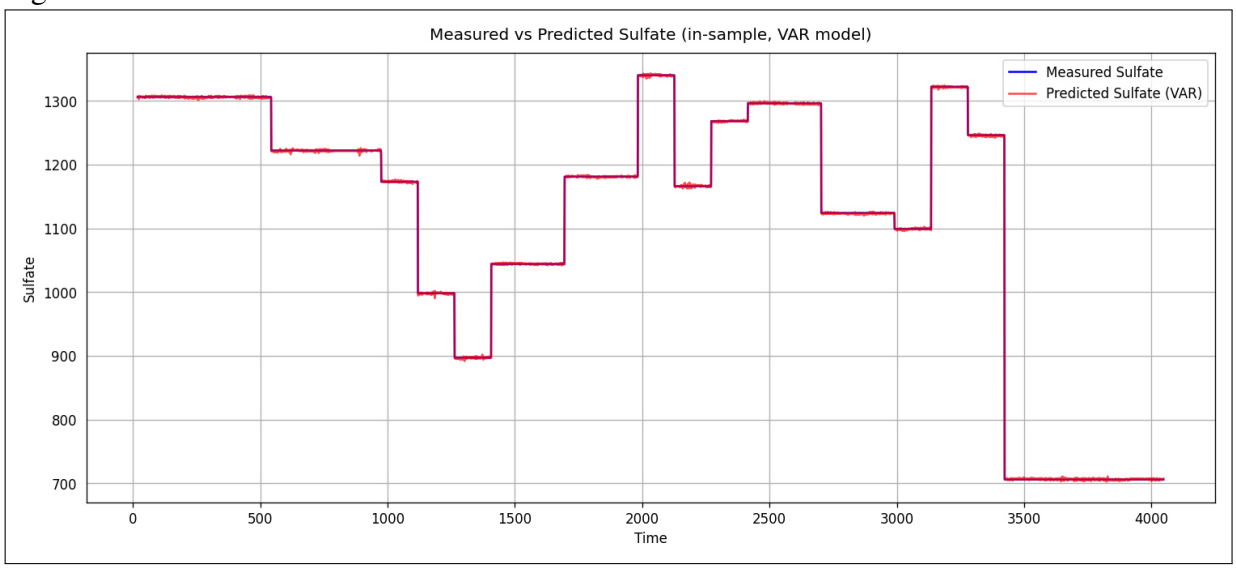
Nevertheless, these findings suggest that a data-driven approach to dissolved oxygen modelling improves process understanding, making it a valuable tool for operators and engineers seeking to optimize aeration energy consumption without compromising effluent quality.

4.2.2 Sulfate prediction

The VAR model for sulfate prediction was first estimated using a lag length of 18, selected based on AIC and FPE minimization criteria. However, an in-depth examination of the equation for sulfate concentration reveals that only a handful of lagged variables are statistically significant, notably Lag 1 of sulfate concentration ($p < 0.001$) and select lags of effluent flow (lags 2, 6, and 7). The extremely high coefficient for Lag 1 sulfate concentration is highlights a high degree of temporal persistence in sulfate concentration, but the low-sampling of this variable cautions for statistical irrelevance. Other lagged values of sulfate concentration exhibit coefficients close to zero and are not statistically significant. Similarly, most of the exogenous operational variables do not show statistically significant effects on sulfate levels across the majority of lags, except for the effluent flow rate at lags 1, 2, 6, and 7, which may indicate a potential, though unstable, relationship with sulfate dynamics that deserves further investigation. Despite reducing the lag order of the VAR model to more parsimonious levels (3 and 7 lags, as suggested by BIC and HQIC), signs of overfitting persisted in the sulfate equation. In each case, the model continued to estimate a large number of parameters relative to the number of variables that were statistically or theoretically meaningful. Most lagged variables remained statistically insignificant ($p > 0.1$) and contributed little to model fit, indicating that the model was still partially capturing random fluctuations or noise in the data rather than true underlying relationships. Furthermore, the dominant role of sulfate's own first lag consistently overshadowed the explanatory power of other inputs, suggesting a largely autoregressive structure. These results suggested that the dynamics of sulfate concentration are predominantly driven by its own past values, with minimal explanatory power contributed by other variables in the system. 20.

To complement the multivariate time series analysis, a series linear regression models were evaluated to assess their ability to capture sulfate dynamics without overfitting due

Figure 20 – VAR model for the sulfate concentration



Source: made by the author (2025).

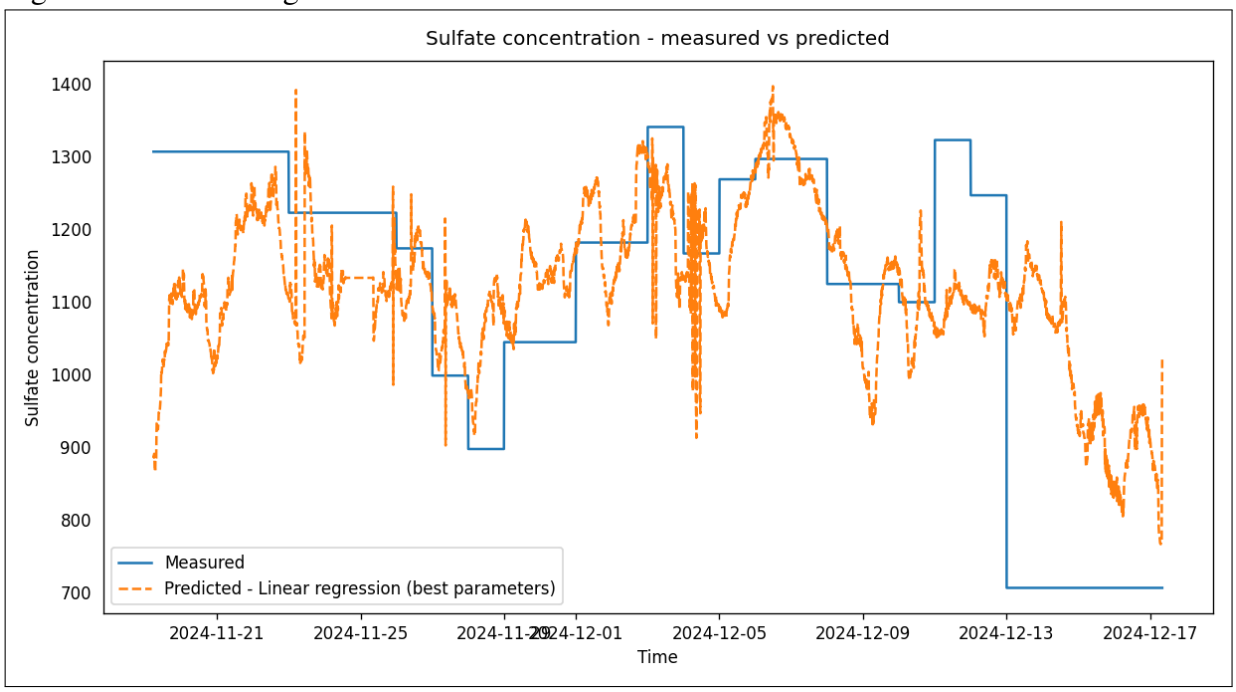
Table 11 – Model performance for sulfate prediction (best hyperparameters)

Model	**MAE**	**RMSE**	R^2
VAR	8.512	10.150	
Linear Regression	127.315	193.831	0.361
Ridge Regression (λ=1000)	128.344	193.407	0.354
LASSO Regression (λ=0.01)	127.311	193.825	0.361

to temporal data. However, even with linear regression models, sulfate concentration proved much more difficult to model using the available dataset (Table 11). The relatively lower R^2 values across all models suggest that sulfate concentration is less predictable from the available sensor data compared to dissolved oxygen. Ridge regression models consistently outperformed Lasso and unregularized linear regression in cross-validation, achieving the lowest RMSE and stable R^2 values. These results indicate that Ridge regression successfully balances fit and generalization, effectively controlling for potential overfitting caused by multicollinearity or high dimensionality in the predictor set. In contrast, some Lasso models excessively penalized the coefficients, leading to underfitting and a higher loss of predictive power. Although standard linear regression achieved similar performance on training data, its lack of regularization makes it more susceptible to capturing noise rather than meaningful patterns. Overall, these findings support the conclusion that regularized linear models, particularly Ridge regression (Figure 21) offer a more accurate and robust representation of sulfate behavior, especially when contrasted with the tendency of high-lag VAR models to overfit.

The results confirm the hypothesis established in the correlation and causality anal-

Figure 21 – Linear regression model for the sulfate concentration



Source: made by the author (2025).

ysis, that sulfate concentrations behave are likely influenced by external industrial discharges or slow-reacting biological processes not captured in the dataset. Additionally, the lack of daily samples during the time window selected for the study made it more difficult to study the autocorrelation for this variable. Therefore, a wider window, with a higher sampling frequency is needed for further studies.

4.2.3 Models' Limitations and Practical Constraints

Overall, the limitations observed in model performance can be attributed to both process complexity and data quality constraints. Given that the dataset lacked enough chemical dosing records, models could only rely on downstream or indirect signals, which failed to explain sulfate variability in a meaningful way, highlighting a core limitation in many industrial WWTPs regarding data availability. Additionally, sensor failures and missing values impaired long-term DO forecasting. These issues reinforce the need for improved data governance, structured operator training, and gradual implementation of automated monitoring systems, as recommended by Bahramian *et al.* (2023).

Despite the constraints, however, the DO forecasting models offered valuable short-term predictions that could inform proactive aeration adjustments, improving process stability and energy efficiency, as suggested by Afan *et al.* (2024). Nevertheless, the findings indicate

that the full operational potential of data-driven models will only be realized with improved data acquisition practices and comprehensive operator training programs, as reliable data and informed operational strategies are prerequisites for better handling of data-driven systems in industrial wastewater treatment operations.

5 CONCLUSION AND FUTURE WORKS

This study examined how data-driven modeling techniques can be applied to forecast key variables in an industrial activated sludge wastewater treatment plant. The models were tested using real operational data, which included both continuous sensor measurements and manually recorded plant operations. Significant challenges were encountered with the data, such as missing values, inconsistent timestamps, and incomplete operational records. These problems directly impacted model calibration and validation, emphasizing the need for structured data management and better operator training to support reliable analytics in such environments.

However, even though the plant already has a functional sensor infrastructure in place, including measurements of variables like dissolved oxygen, temperature, and pH, the full potential of this infrastructure is not being realized, mostly due to infrequent calibration and inconsistencies in data recording. Improving the regularity of sensor calibration and ensuring systematic data collection should significantly enhance the quality of information available for modeling. Similarly, although laboratory analyses are routinely performed, their low temporal resolution limits their usefulness for forecasting. More frequent lab measurements, particularly of sulfate concentrations, may improve the models' ability to capture variability, thus enabling the possibility of supporting process control.

The results showed that forecasting dissolved oxygen levels was feasible during periods of operational stability, which could enable more proactive aeration control, helping to stabilize the process and improve energy efficiency. In fact, DO forecasting stands out as a practical application ready for real-time use. It offers tangible operational benefits, such as better aeration control and energy savings, and can serve as an accessible starting point for adopting data-driven methods in plant operations. On the other hand, predicting sulfate concentrations was proven to be more difficult, given the low frequency of laboratory measurements and irregular patterns of industrial discharges, which made it hard for the models to detect and follow the dynamic behavior of this variable. This difficulty is consistent with challenges described in the literature and highlights the importance of high-frequency data collection for accurate forecasting in industrial systems.

The findings in this work suggest that time series and statistical learning models can improve process monitoring in industrial wastewater treatment plants, but their success depends heavily on the quality of data, the analytical tools available, and the skills of the operational team. Improving how data is collected and recorded, training operators to work with sensors and

analytics tools, and gradually incorporating process knowledge into the models can make these approaches more effective and easier to interpret. The operational team should aim to improve data infrastructure and improve the frequency and regularity of laboratory sampling, in order to enable the enhancement of model performance and their applicability at the plant.

Ultimately, this study reinforces the idea that, while data-driven models offer valuable opportunities for optimization in industrial wastewater treatment, their practical value relies on dependable data acquisition, consistent operational practices, and the development of a data-oriented culture within the plant. Indeed, success in this area requires a combination of statistical modeling, engineering knowledge, and real-world process insight.

REFERENCES

AFAN, H. A.; MOHTAR, W. H. M. W.; KHALEEL, F.; KAMEL, A. H.; MANSOOR, S. S.; ALSULTANI, R.; AHMED, A. N.; SHERIF, M.; EL-SHAFIE, A. Data-driven water quality prediction for wastewater treatment plants. **Heliyon**, v. 10, p. e36940, 2024.

BAHRAMIAN, M.; DERELI, R. K.; ZHAO, W.; GIBERTI, M.; CASEY, E. Data to intelligence: The role of data-driven models in wastewater treatment. **Expert Systems with Applications**, v. 217, p. 119453, 2023.

BATINI, M. S. a. C. **Data and Information Quality: Dimensions, Principles and Techniques**. 1. ed. [S.l.]: Springer, 2016. (Data-Centric Systems and Applications). ISBN 9783319241043; 3319241044; 9783319241067; 3319241060.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time Series Analysis: Forecasting and Control**. 5th. ed. [S.l.]: John Wiley & Sons, 2015.

COSENZA, B.; CONCAS, A.; PICONE, A.; MESSINEO, A.; De Blasi, V.; SCALICI, B.; VOLPE, M. A data-knowledge hybrid decision support system for wastewater treatment operations: The acqua dei corsari plant case study. **Information Sciences**, v. 718, 2025.

DÜRRENMATT, D. J.; GUJER, W. Data-driven modeling approaches to support wastewater treatment plant operation. **Environmental Modelling & Software**, v. 30, p. 47–56, 2012.

ELSAYED, A.; SIAM, A.; EL-DAKHAKHNI, W. Machine learning classification algorithms for inadequate wastewater treatment risk mitigation. **Process Safety and Environmental Protection**, v. 159, p. 1224–1235, 2022.

GUJARATI, D.; PORTER, D. **Basic Econometrics**. 5th ed.. ed. New York: McGraw Hill Inc., 2009.

HAIMI, H.; MULAS, M.; CORONA, F.; VAHALA, R. Data-derived soft-sensors for biological wastewater treatment plants: An overview. **Environmental Modelling Software**, v. 47, p. 88–107, 2013.

HASTIE, T.; TIBSHIRANI, R. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd ed.. ed. New York: Springer, 2009.

HENZE, M.; GUJER, W.; MINO, T.; LOOSEDRECHT, M. van. **Activated Sludge Models ASM1, ASM2, ASM2d and ASM3**. [S.l.]: IWA Publishing, 2006.

HYNDMAN, R.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. 3rd. ed. Australia: OTexts, 2021.

JENKINS, D.; WANNER, J. **Activated Sludge – 100 Years and Counting**. [S.l.]: IWA Publishing, 2014.

KHATRI, V.; BROWN, C. V. Designing data governance. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 1, p. 148–152, Jan. 2010. ISSN 0001-0782.

MACKINNON, J. G. Numerical distribution functions for unit root and cointegration tests. **Journal of Applied Econometrics**, p. 601–618, 1996.

MEDEIROS, G. J. de; FARIAS, F. P. de; SOUSA, D. R. de; MULAS, M. **Overview of ICA in Brazilian WWTPs.** [S.l.]: 14th IWA International Conference on Instrumentation, control and Automation, 2025.

Metcalf & Eddy Inc.; TCHOBANOGLOUS, G.; STENSEL, H. D. **Wastewater Engineering: Treatment and Resource Recovery**. 5th. ed. New York: McGraw-Hill Education, 2014.

NEWHART, K. B.; HOLLOWAY, R. W.; HERING, A. S.; CATH, T. Y. Data-driven performance analyses of wastewater treatment plants: A review. **Water Research**, v. 157, p. 498–513, 2019.

ORHON, D.; BABUNA, F.; KARAHAN, O. **Industrial Wastewater Treatment by Activated Sludge**. [S.l.]: IWA Publishing, 2009. ISBN 9781843391449.

QUAN, Y.; HAN, H.; ZHENG, S. Effect of dissolved oxygen concentration (microaerobic and aerobic) on selective enrichment culture for bioaugmentation of acidic industrial wastewater. **Bioresource Technology**, v. 120, p. 1–5, 2012. ISSN 0960-8524.

RIEGER, L.; GILLOT, S.; LANGERGRABER, G.; OHTSUKI, T.; SHAW, A.; TAKACS, I.; WINKLER, S.; RIEGER, L. **Guidelines for using activated sludge models**. [S.l.]: IWA Publishing, 2012. 312 p.

SHUMWAY, R. H.; STOFFER, D. S. **Time Series Analysis and Its Applications**. Berlin, Heidelberg: Springer-Verlag, 2005. ISBN 0387989501.

SPERLING, M. von; VERBYLA, M. E.; OLIVEIRA, S. M. A. C. **Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners**. [S.l.]: IWA Publishing, 2020.

SUN, A. Y.; SCANLON, B. R. How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. **Environmental Research Letters**, IOP Publishing, v. 14, n. 7, p. 073001, jul 2019.

TCHOBANOGLOUS, G.; BURTON, F. L.; STENSEL, H. D. **Wastewater Engineering: Treatment and Reuse**. 4th. ed. New York: McGraw-Hill Education, 2003.

WANG, R.; PAN, Z.; CHEN, Y.; TAN, Z.; ZHANG, J. Influent quality and quantity prediction in wastewater treatment plant: Model construction and evaluation. **Polish Journal of Environmental Studies**, v. 30, n. 5, p. 4267–4276, 2021. ISSN 1230-1485.

XUE, W.; HAO, T.; MACKEY, H. R.; LI, X.; CHAN, R. C.; CHEN, G. The role of sulfate in aerobic granular sludge process for emerging sulfate-laden wastewater treatment. **Water Research**, v. 124, p. 513–520, 2017. ISSN 0043-1354. Available at: <https://www.sciencedirect.com/science/article/pii/S0043135417306619>.