ACIT5900

MASTER THESIS

in

Applied Computer and Information Technology (ACIT)
May 2023

Robotics and Control

Prediction of influent composition in wastewater and sludge based on Statistical and Machine Learning models.

Bipasha Mukherjee

S366272

Department of Mechanical, Electronics and Chemical Engineering

Faculty of Technology, Art and Design

OSLOMET

Acknowledgement

This project report is part of the course ACIT5900-1 21H Master thesis(short) Research. This report includes research on influent composition prediction of wastewater and sludge treatment process using statistical model for time series analysis and machine learning models. Being a student of Robotics and Control system, I can use my knowledge of both process control and automation in this thesis topic. This makes it easy to choose the thesis topic in this field.
I want to express my sincere gratitude to my professors Tine Komulainen, Olga Korostynska, Arvind Keprate, Simen Antonsen, Rafael Borrajo, Per Ola Rønning, and Yasha Parvini for giving me valuable guidance through all master thesis discussion sessions. Regular weekly meeting sessions and presentations helped me to proceed further. They always helped me whenever or wherever I faced difficulties in the whole journey of master thesis. They also support in getting additional resources and facilities from OsloMet.
I would like to thank Hias Water Resource Recovery Facility and VEAS Water Resource Recovery Facility for providing the datasets and helping to work with their process data. I would also like to thank OsloMet for giving me opportunity to study master in the university.
I would also like to express my sincere gratitude to my classmates , Einar Nermo, Mohamed Abdishakur Mohamed, Bilal Mukhter to explain all my doubts through discussions, provide the support to complete the thesis assignments and thesis work .
More personally, I want to thank my family for their patience, motivation, and support during my study and without whom this would not have been possible.

Bipasha Mukherjee
Oslo, Norway
May 2023

Abstract

For the optimized operation of a wastewater resource recycle facility (WWRF), it is essential to consider significant disturbances such as fluctuations in the influent flow rate and wastewater compositions. The online monitoring of influent characteristics is limited by scarce instrumentation and high costs. This study demonstrated influent composition prediction of two different wastewater treatment plants (WWTPs), with wastewater and sludge treatment process. Data-driven models (statistical models used for time series analysis/ Machine learning model) have been developed using HIAS wastewater treatment process and VEAS sludge treatment process data to predict the influent compositions.

In this work, statistical models for time series analysis such as ARIMA (Autoregressive Integrated Moving Average) and SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous input), Linear regression, Lasso, Ridge regression and different machine learning algorithms such as Random Forest (RF), Decision Tree (DT), Support Vector Regression (SVR) and Artificial Neural Network (ANN) were examined and compared. These models were developed to detect inlet phosphate (PO_4), and inlet soluble chemical oxygen demand(sCOD) in wastewater inlet organic acid in sludge, which served as output variables.

In both processes, Linear regression, Ridge regression and Neural Network consistently demonstrated the best performance for evaluation estimation as evidenced by the lowest values of Root Mean Square Error (RMSE), and the highest coefficient of determination (R^2). SARIMAX exhibited acceptable results with R^2 as 0.95 in organic acid prediction modeling. In contrast, ARIMA and SARIMAX algorithms in Hias datasets did not meet the requirements because of the complex and nonlinear structure of the dataset issue. This study offers an efficient method for forecasting the quality of wastewater and sludge influents, which can be advantageous for process control and thereby contributing to the zero-pollution objective of the European Green Deal and to the European Missions such as one on 'Restoring our ocean and waters by 2030'.

Keywords: Wastewater treatment, Sludge treatment, influent composition, SARIMAX, Multivariate Regression, Artificial Neural Network

Contents

List of symbols and abbreviations

WWTP Wastewater treatment plants

sCOD Soluble Chemical oxygen demand (mg/L)

BOD Biochemical oxygen demand (mg/L)

PO_4 Phosphate(mg/l)

NH_4-N Ammonia(mg/l)

NO_2 Nitrate(mg/l)

NO_3 Nitrite(mg/l)

TS Total suspended solid

ANFIS Adaptive neuro-fuzzy inference system

ARIMA Autoregressive integrated moving average

SARIMA Seasonal autoregressive integrated moving average

SARIMAX Seasonal autoregressive integrated moving average with exogenous input

NN Neural network

BP Backpropagation

DL Deep learning

DNN Deep neural network

DT Decision tree

ML Machine learning

LSTM Long short-term memory

SVM Support vector machines

SVR Support vector regression

RF Random Forest

List of figures

List of tables

1. Introduction

1.1 Background

Severe water pollution caused by rapid population growth and industrialization has threatened the environment and human society. Variations in the quality and quantity of wastewater at the inlet to wastewater treatment plants (WWTP) often reduce treatment efficiency. As a result, WWTP can discharge effluent with above-limit compositions like Chemical oxygen demand (COD), Biological oxygen demand (BOD), Total nitrogen (TN), and Phosphate (PO_4) concentrations. (Andreides et al., 2022)

One of the crucial strategies to restrict those pollutants into the environment is to optimize the efficiency of WWTP (Ly et al., 2022). The accurate forecast of wastewater treatment plant (WWTP) key features can be advantageous for plant in several ways:

- **To support process design and controls:** The influent composition prediction in advance can help the plant operators adjust and optimize the treatment processes and improve treatment efficiency.
- **Reduce operational costs:** Optimization of treatment process, in terms also help to reduce energy and chemical consumption, leading to cost savings for the plant.
- **Improve system reliability**: Influent composition prediction can help to identify potential issues and challenges that may affect the treatment process such as high organic load, nutrient imbalances in advance which help to improve system reliability.
- **Predictive Maintenance:** Accurate influent composition prediction can help plant operators anticipate maintenance needs, such as cleaning filters, unclogging pipes, or replacing equipment. This can help reduce downtime and maintenance costs. (Cheng et al., 2020).

Influent wastewater compositions vary depending on the source, region, and dry or wet and warm or cold season. Wastewater from different industries like papermaking, petrochemical, and food processing are typically enriched in organic carbon and have a low nutrient content, e.g.,

nitrogen and phosphorus. In contrast, domestic wastewater contains high amounts of nutrients. In municipal wastewater chemical composition of influent wastewater could also vary in ratio mixtures (e.g., between households, industries, and surface runoff), sewer system length and type (i.e., separate or combined), and transformation processes within sewerage networks. Apart from this, wastewater quality varies with human behavior-dependent features based on the season, day of the week, and time of the day. (Ly et al., 2022). Wastewater flow is relatively stable during dry periods whereas flow rate and compositions changes originating from intensive precipitation (e.g., heavy rainfall / snow melt) hugely affect the correct prediction of wastewater quantity and quality. (Andreides et al., 2022)

It is essential to observe the quantity and quality of various inlet wastewater and sludge parameters to remove pollutants, maintain effluent quality, and reduce energy consumption during the plant-scale treatment processes. Offline measurements, hardware sensors, or recently highlighted model-based soft sensors monitor those parameters, generating a voluminous amount of environmental multivariate time-series data. Selecting a suitable mathematical model to accurately predict wastewater and sludge treatment is challenging due to their complex composition, different treatment mechanisms, and environmental data's non-linear, dynamic, and periodic nature. (Cheng et al., 2020) (Andreides et al., 2022)

Various mathematical models and machine learning algorithms for forecasting water-related variables have recently been developed and even calibrated with full-scale historical data. Model-based mathematical approaches and data-based machine -learning algorithms have fewer assumptions, capable of analyzing vast datasets with less processing preparation and computing time. Thus, they effectively handle the complex nonlinear, unstable, and interdisciplinary features of water quality parameters. (Ly et al., 2022) (Andreides et al., 2022). Some important types of time-series mathematical model to identifying the highly nonlinear systems are, Auto-Regressive with eXternal model input (ARX), Auto Regressive moving Average with eXternal model input (Armax), Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) Partial Least Squares, Lasso, Ridge. Linear time-series models such as Autoregressive moving-average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) with the ARIMA model being the most

widely used one. However, it should be noted that the assumption of data input for the linear and normal time series distribution does not typically hold in WWTPs, especially for the real-time monitoring sensors, of which biofilm formation, precipitates, and solid deposition could easily influence the measurement accuracy. Consequently, this could result in more missing data, outliers, uncertainty, and reduced overall prediction performance. In comparison to this machine learning models such as Deep Neural Networks (DNNs), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), fuzzy-based model, i.e., Adaptive Network based Fuzzy Inference System (ANFIS), could offer more advanced functions to reveal nonlinear properties of wastewater variables (Ly et al., 2022).

Research questions

1. Which prediction model (statistical or machine learning) is more beneficial to accurately forecast influent composition in HIAS-process?
2. Possibility of sludge quality prediction of VEAS-sludge treatment process using the same prediction model as WWTP?

1.2 Theoretical Background

The real-time industrial data for influent compositions of wastewater and sludge treatment process is complex. To design, test, and compare accuracy of different data-driven mathematical and machine learning models, it is important to first understand the theoretical and mathematical background of different models.

1.2.1 ARIMA model

ARIMA (Autoregressive integrated moving average) model is one of the most widely used approaches to time series forecasting. ARIMA models aim to describe the autocorrelations in the data. The ARIMA model was developed by Box and Jenkins in 1976. An ARIMA model consists of three components, which are the autoregressive (AR) process, the moving average (MA) process, and the integrated component (I).

Autoregressive Component — AR(p): In an autoregression model, the predicted variable uses a linear combination of *past values of the same variable*. The term *auto*regression indicates that it

is a regression of the variable against itself. AR model of order p assumes that each observation (Y_t) is a linear combination of prior observations ($Y_{t-1}, Y_{t-2,\ldots\ldots}Y_{t-p}$) and a random error component (ε_t). The AR equation can be written as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \qquad (1.1)$$

where ϕ1 . . . ϕp are the AR model parameters and ε_t is white noise. The highest value of p, for which ϕp $\neq$ 0, is the order of the AR process. Thus, the AR process of order p can be denoted by AR(p).(Rob J Hyndman and George Athanasopoulos, n.d.)(Zhang et al., 2019)

Moving average Component — MA(q):

Rather than using past values of the predicted variable in a regression, a moving average model uses past predicted errors in a regression-like model. The MA process of order q assumes that each observation (Y_t) is a linear combination of prior error components, ($\varepsilon_{t-1}, \varepsilon_{t-2,\ldots\ldots}\varepsilon_{t-q}$) and a random error component (ε_t). The MA equation can be written as:

$$Y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (1.2)$$

where θ1 . . . θq are the MA model parameters and ε_t is white noise. The highest value of q, for which θq $\neq$ 0, is the order of the MA process. Thus, the MA process of order q can be denoted by MA(q).(Zhang et al., 2019) (Brendan Artley, 2022)

Integrated component (I): The I component is for addressing the non-stationarity of the time series. A time series is stationary if its statistical properties (e.g., mean, variance, autocorrelation) are constant over time. Since the data series in ARIMA needs to be stationary, a non-stationary time series needs to be transformed through differencing. Differencing eliminates the trend and seasonality of time series data by calculating the difference between consecutive observations such as:

$$Y_t' = Y_t - Y_{t-1} \qquad (1.3)$$

Sometimes the data should be differenced twice to obtain stationarity, which is known as second order differencing like:

$$Y_t'' = Y'_t - Y'_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \qquad (1.4)$$

The number of transformations that are required to obtain stationarity is denoted by d. A process which has dth order differencing transformations is called an integrated process of order d, and it can be denoted as I(d).(Zhang et al., 2019)

If we combine difference with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for Autoregressive Integrated Moving Average ("integration" is the reverse of differencing). The full model can be written as

$$Y'_t = \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + \cdots + \phi_p Y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (1.5)$$

Where Y'_t is the differenced series. The "predictors" on the right-hand side include both lagged values of Y_t and lagged errors. (Rob J Hyndman and George Athanasopoulos, n.d.)Thus, an ARIMA model with the order of (p, d, q) is a combination of AR(p), MA(q), and I(d), where,

p= order of the autoregressive part; d= degree of first differencing involved; q= order of the moving average part.

1.2.2 SARIMA Model

The ARIMA model is good, but it can only handle non-seasonal data. To include seasonality and exogenous variables in the model can be extremely powerful. Since the ARIMA model assumes that the time series is stationary, we need to use a different model. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models and it is written as follows:

SARIMA	(p,d,q)	(P,D,Q)s
	↑↑	↑↑
	Non-seasonal part	Seasonal part of
	of the model	of the model

where s= frequency of seasonality.

$$Y'_t = \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + \cdots + \phi_p Y'_{t-p} \quad + \quad \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \alpha_1 Y'_{t-1s} + \alpha_2 Y'_{t-2s} + \cdots + \alpha_P Y'_{t-P+} + \eta_1 \varepsilon_{t-1s} + \eta_2 \varepsilon_{t-2s} + \cdots + \eta_Q \varepsilon_{t-Qs} + \varepsilon_t$$

$$Y'_t = \sum_{n=1}^{p} \phi_n Y'_{t-n} + \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{n=1}^{P} \alpha_n Y'_{t-sn} + \sum_{n=1}^{Q} \eta_n \varepsilon_{t-sn} + \varepsilon_t \qquad (1.6)$$

This model is very similar to the ARIMA model, except that there is an additional set of autoregressive and moving average components. The additional lags are offset by the frequency of seasonality (ex. S=12 for monthly, s=24 for hourly, s=7 for daily).

SARIMA models allow for differencing data by seasonal frequency, yet also by non-seasonal differencing. Knowing which parameters are best can be made easier through automatic parameter search frameworks such as pmdarima.(Brendan Artley, 2022; Rob J Hyndman and George Athanasopoulos, n.d.)

1.2.3 SARIMAX

SARIMAX (Seasonal Autoregressive Moving Average with Exogenous Input) is a popular time series model that combines SARIMA model with exogenous variables. It can be used to forecast future values of a time series based on its own past values as well as the past values of one or more exogenous variables.

Exogenous variables are also called covariates and can be thought of as parallel input sequences that have observations at the same time steps as the original series. The primary series may be referred to as endogenous data to contrast it from the exogenous sequence(s). The observations for exogenous variables are included in the model directly at each time step and are not modeled in the same way as the primary endogenous sequence (e.g. as an AR, MA, etc. process). SARIMAX model can be written as,

$$Y'_t = \sum_{n=1}^{p} \phi_n Y'_{t-n} + \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{n=1}^{r} \beta_n x_{n_t} + \sum_{n=1}^{P} \alpha_n Y'_{t-sn} + \sum_{n=1}^{Q} \eta_n \varepsilon_{t-sn} + \varepsilon_t \quad (1.7)$$

This model considers exogenous variables.(Brendan Artley, 2022)

1.2.4 Regression

Regression: Regression is a statistical approach for estimating the relationship between the dependent variables and one or more independent variables or predictors. Regression analysis is generally used when dealing with a dataset with the target variable in the form of continuous data. Regression analysis explains the changes in output with changes in selected predictors. The expected outcome depends on the predictors, which determine the average value of the dependent variables when the independent variables are altered. Regression analysis serves

three primary purposes: evaluating the power of predictors, projecting an outcome, and predicting trends. Regression analysis relies on statistics and can provide dependable outcomes for identifying both linear and non-linear correlations between independent and dependent/target variables.(Veena Ghorakavi, n.d.-a)

There are several types of regression techniques, each suited for different types of data and different types of relationships. The main types of regression techniques are:

1. Linear Regression
2. Polynomial Regression
3. Stepwise Regression
4. Decision Tree Regression
5. Random Forest Regression
6. Support Vector Regression
7. Ridge Regression
8. Lasso Regression
9. Elastic Net Regression
10. Bayesian Linear Regression

Before going deeper into different regression processes, it is important to discuss some necessary terms related to regression analysis in machine learning.

Overfitting is a phenomenon that occurs when a regression model is constrained to the training set and not able to perform well on unseen data. This is the case when the model memorizes the training data instead of learning the patterns in it.

Underfitting is the case when model is not able to learn even the basic patterns available in the dataset. The underfitted model is unable to perform well on both the training data and validation data. In this case increase the complexity of the model or add more features to the feature set.

Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.

Linear Regression

Linear regression is a straight line that fits a series of points on a two-dimensional plane and is used for predictive analysis. It is a linear approach for modeling the relationship between the criterion or the output response and the multiple predictors or input variables. *Linear regression* is a method that examines the conditional probability distribution or trend of training samples to forecast new sample points.

The general formula for linear regression is:

y = Ax + B

where, A is the model weights or parameters, B is known as the bias.

After training the model using data from the training set, the optimal values of the two parameters A and B in the equation can be determined and used to predict newly observed samples to obtain the predicted value of y. Linear regression is a fundamental type of regression analysis that can model a linear connection between one dependent variable and one or more independent variables.

This method is called linear regression because the model is composed of linear combinations of all features and can be written as:

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n + \varepsilon. \quad (1.8)$$

$\hat{y}$: Represents the predicted value of a dependent variable

n: Represents the number of features

x_n: Represents the observation of the nth feature

θ_n: Represents the value of the nth parameter

$\theta_n x_n$: Represents the regression coefficient of the nth independent variable

ε : model error (how much variation there is in estimation, cost function

it is also necessary in regression analysis to define the appropriate cost function, which quantifies the error between the predicted and observed values. After selecting an appropriate cost function, the training process identifies the minimum value. For the linear regression algorithm, the most commonly used cost function is the MSE function and R^2. (Nair et al., 2022)(Veena Ghorakavi, n.d.-b)(Wang et al., 2021)

Polynomial Regression

Polynomial regression is an extension of linear regression and is used to model a nonlinear relationship between the dependent and independent variables. In many real-time cases where the equation of the line does not fit the data well, polynomial regression may be an alternative. In polynomial regression, syntax remains the same as linear regression but includes higher powers (such as square or cubic terms) in the input variables. It helps to increase the model freedom and to capture nonlinear changes in the data. Including polynomial terms in a model can make it more complex, but also increase its capacity to fit data. However, this may lead to a higher risk of overfitting, despite reducing the training error. In polynomial regression, the most important parameter is the degree of the highest power. If the degree of the highest power is n and there is only one characteristic, the polynomial regression equation can be expressed as:

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \ldots \theta_{n-1} x^{n-1} + \theta_n x^n + \varepsilon. \qquad (1.9)$$ (Wang et al., 2021)

Stepwise regression

Stepwise regression is used for fitting regression models with predictive models. It is carried out automatically. With each step, the variable is added or subtracted from the set of explanatory variables. The approaches for stepwise regression are forward selection, backward elimination, and bidirectional elimination.(Veena Ghorakavi, n.d.-a)

Decision Tree Regression

A Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute,

each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. There is a non-parametric method used to model a decision tree to predict a continuous outcome.

Random Forest Regression

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

Support Vector Regression (SVR)

Support vector regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value.

SVR can use both linear and non-linear kernels. A linear kernel is a simple dot product between two input vectors, while a non-linear kernel is a more complex function that can capture more intricate patterns in the data. The choice of kernel depends on the data's characteristics and the task's complexity.(Veena Ghorakavi, n.d.-a)

Lasso regression

Lasso regression is a technique for regression analysis that accomplishes variable selection and regularization simultaneously. It implements L1 Regularization and is referred to as LASSO (Least Absolute Shrinkage and Selection Operator) regression.

Lasso Regression involves adding a penalty term to the loss function (L) that considers the absolute value of the coefficient's magnitude. This method assists in feature selection by penalizing weights that are not useful to the model and bringing them closer to zero.

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{m}|w_i| \qquad (1.10)$$

where, m =Number of Features, n = Number of Examples, y_i =Actual Target Value,

$\hat{y}_i$ =Predicted Target Value

Ridge regression

Ridge regression is a method used to analyze data with multiple regression. In situations where multicollinearity is present, the least squares estimates remain unbiased.

When using polynomial regression, if the polynomial's highest degree is large, the model is at risk of overfitting. Therefore, regularization is frequently used to address this issue. Ridge Regression, also known as L2 regularization, is a method to prevent overfitting during linear regression. The only difference between ridge regression and polynomial regression is the cost function. (Wang et al., 2021) Ridge regression adds "*squared magnitude*" of the coefficient as a penalty term to the loss function(L).The cost function of ridge regression is shown as:

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{m}w_i^2 \qquad (1.11)$$

Elastic Net Regression

When there are many features in the dataset, some of which are irrelevant to the predictive model, it makes the model more complex with a too-inaccurate prediction on the test set (or overfitting). Such a model with high variance does not generalize on the new data. So, to deal with these issues, Elastic net regression combines both L-2 and L-1 regularization to get the benefits of both Ridge and Lasso simultaneously. The resultant model has better predictive power With the help of an extra hyperparameter that controls the ratio of the L1 and L2 regularization.

It performs feature selection and makes the hypothesis simpler. The modified cost function for Elastic-Net Regression is :

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left((1-\alpha)\sum_{i=1}^{m}|w_i| + \alpha\sum_{i=1}^{m}{w_i}^2\right) \qquad (1.12)$$

Where, $\lambda(1-\alpha)$ is the regularization strength for the L1 norm**.**

$\lambda\alpha$ is the regularization strength for the L2 norm.(Wang et al., 2021)(AlindGupta, n.d.)(Veena Ghorakavi, n.d.-a)

1.2.5 Neural Network

ANN is a machine learning technique that could be used to build predictive models by extracting information from past experience. It shows significant improvements of model flexibility compared to traditional predictive models.(Zhang et al., 2019)

An Artificial Neural Network (ANN), or called Neural Network (NN) system, is a computing system inspired by the biological neural networks. An ANN is based on a collection of an interconnected set of several simple computational elements called 'artificial neurons'. The network is built of different layers, consisting of different types of neurons connected with the previous layer. In a general neural network, there are mainly three types of layers:

• Input layer consists of different types of inputs that will come as a part of attributes based on which the desired output will be achieved from the network.

• Hidden layer unit, which consists of multiple neurons. These layers are mainly used as high-dimension to lower-dimension data transfer.

• Output layer, which consists of output neurons that will give the desired output based on inputs.

A simple 1-hidden layer neural network has been described to demonstrate how this algorithm works. The input layer is in the form of a matrix (x) or vector, which will be then multiplied with randomized initialized weights (W). A bias term will be added with this multiplication data, which will be passed through a non-linear function to achieve the next hidden layer output. The size of the weights initialization will be of

$$n_h * n_x \qquad (1.13)$$

where n_h, is the number of neurons in the hidden layer and n_x, is the number of neurons in input layers. The equation of the first hidden layer is

$$h = A(W \cdot x + b_1) \qquad (1.14)$$

where A is the Activation Function. The dot multiplication is the element-wise multiplication here. Based on the value of the hidden layer, the output value will be produced. Like the hidden layer, the weights will multiply with the hidden value as, in this case, hidden value is the input, and with that, the bias will be added. Then the result will be passed through a non-linear function that gives the desired output function.

$$y = A(W1 \cdot h + b_2) \qquad (1.15)$$

where y is the output produced by the neural net. The size of the weights which is coming to the output neuron from the hidden layer is of size

$$n_h * n_y \qquad (1.16)$$

where n_y is the size of the output neuron.

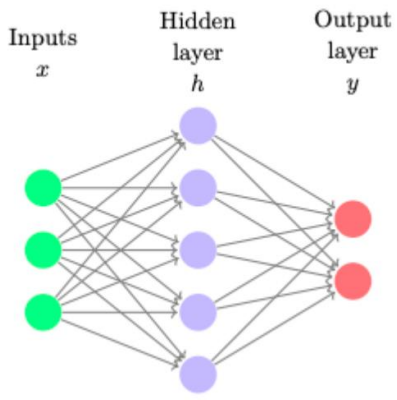


Figure 1.1: Example of Neural Network with One Hidden Layer

In the neural network, the first run is with randomly initialized weights and then updated to minimize the prediction error through an optimization process. There are mainly two steps in a neural network: -

• **Feedforward**, which means that with the weights, the output will be derived based on the input and then hidden layers outputs and the non-linear functions. The flow is left to right, which means input layers to output layers. The learning process of a feedforward network occurs in the perceptron, which consists of a single neuron with adjustable weights and an

activation function.(Zhang et al., 2019)

• **Backpropagation** means that after the output has been derived, it will be compared to the standard output and calculated the loss. Once the loss has been calculated, the error of the weights will be distributed going backwards from output to input direction to achieve the standardized weights.

The above steps are done in one iteration, called one **epoch**. Several epochs can be based on the stabilized and minimum loss on the output.

To achieve the best weights and minimize the loss, **gradient descent algorithm** has been formulated. This gradient descent algorithm is performed with the help of the backpropagation algorithm. The gradient descent algorithm is nothing, but the partial derivative of the errors based on the activation function in each layer. The weights and biases are then adjusted based on this error calculation. In a simple case, if we take $I_s(f)$ as the error calculated in the output layer, then the below process will happen to adjust the weights and biases in each iteration.

$$W = W - \alpha \delta W \, I_s(f) \qquad (1.17)$$

where W is the weights, α is the learning rate, and it is multiplied by the change of error based on weights. The same will be applied to biases to adjust the biases in each layer.

$$b = b - \alpha \delta b \, I_s(f) \qquad (1.18)$$

The derivation does not happen in the first layer as that is the input layer, and that does not need to be adjusted.

The number of neurons in each layer and the number of hidden layers are called **hyperparameters** of the neural network. To tune the performance of the neural network, the hyperparameters need to be tuned. (Zhang et al., 2019)

K-Fold Cross-Validation for Neural Networks

Cross-validation is a process used to estimate a neural network's quality. By applying cross-validation to multiple neural networks that have varying parameter values (number of hidden nodes, back-propagation learning rate), the results can be analyzed to determine the best set of parameter values.

To implement k-fold cross-validation with a neural network, the data is first divided into k subsets. Then, the neural network is trained on k-1 of the subsets and tested on the remaining subset. This process is repeated k times, with each fold being used as the test set exactly once. The network performance is then evaluated by averaging the results of the k experiments.

K-fold cross-validation can be used for a variety of neural network architectures, including feedforward neural networks, convolutional neural networks, and recurrent neural networks. It is particularly useful when working with limited data since it allows for the efficient use of all available data for training and testing the network.(DR NILIMESH HALDER, 2019)

1.3 Objectives of the thesis

This work aims to design, test, and validate a data-driven mathematical and machine learning model to estimate influent nutrient composition from two different wastewater treatment plant:

1. Inlet composition of wastewater treatment process, such as PO_4, sCOD, and influent flow rate in a full-scale municipal wastewater resource recovery facility in Hamar, Norway.
2. Inlet composition of sludge treatment process, such as Organic Acid, TS% from one of the Norway's biggest wastewater and sludge treatment process in Asker.

The real-time online data with a 10min sampling frequency are compared to the data obtained from periodic (once daily) lab-analyzed influent flow rate and other compositions. A concise data-

analysis tool has been used for cleaning and transforming online and lab-analyzed data and visualizing the relationship of different parameters.

The state-of-the-art algorithm deployment strategy of various time-series mathematical and machine-learning models has been done in this paper. After that, a well-established comparison of accuracy between different models is discussed.

2. Literature Review

To grasp the dynamic changes of influent nutrients composition and flow rate of WWTP in real-time, machine learning, mathematical modelling, and other forecasting methods are increasingly being used to assist WWTP operation and management.

Data driven statistical models and machine learning models used for influent prediction:

2.1. Data-driven models

Data-driven models are usually based on large sets of data (big data). The models aim to find the relationship between state variables and control variables, i.e., predictors and predicted output, with no additional knowledge about the internals of the processes. These models mainly predict selected parameters, detect unusual and faulty situations, and develop soft sensors. Although such modeling seems less popular than mechanistic due to lacking deep insight into the process, data-driven models can be more accurate than mechanistic for influent flow prediction. Moreover, the connection of a data-driven control strategy with artificial intelligence capable of self-learning can detect malfunctioning of the treatment process or sensor. It can thus timely detect impending equipment failure and choose a backup process strategy.(Andreides et al., 2022)

Most popular wastewater influent prediction models can be created using statistical methods, such as linear (ARMA, ARIMA, multiple regression, etc.) or nonlinear (SARIMAX, polynomial, exponential, regression trees, random forests, boosted trees, neural networks, support vectors, etc.) models.(Wodecka et al., 2022)

2.1.1. ARIMA

ARIMA model has been studied mainly for flow rate prediction.(Boyd et al., 2019)(Zhang et al., 2019). In all studies, historical flow rate data with various sampling frequencies were used, while accuracy was expressed by the correlation coefficient and the coefficient of determination(R^2).(Andreides et al., 2022).

(Boyd et al., 2019) used ARIMA for daily influent flow forecasting where five wastewater treatment stations across North America are used to validate ARIMA's performance. Sampling frequency of flow data was 5min, 15min and daily and results demonstrate that ARIMA models can produce satisfactory daily influent flow forecasts with R^2= 0.89.

In Canada, (Zhang et al., 2019) developed ARIMA model for predicting wastewater inflow to address challenges such as precipitation-runoff relationships in combined sewer systems, unpredictability due to aging infrastructure, and frequently inconsistent data quality. Here, fifteen-minute flow data over a period of 1 year were collected, and the resampled daily flow data were used to train and validate the developed models. The model performances were good with R^2 as 0.78 for training data and 0.63 for testing data.

2.1.2. SARIMA

(Do et al., 2022)aims to investigate patterns of the wastewater inflow behavior and develop a seasonal autoregressive integrated moving average (SARIMA) forecasting model at low temporal resolution (hourly) for a short-term period of 7 days for a wastewater treatment facility in South Australia. Historical wastewater inflow data collected for a 32-month period and result shows presence of seasonality with dependence on time of the day and day of the week.

2.1.3. SARIMAX

Due to significant variation of influent wastewater constituents and complex treatment processes within wastewater treatment plants, real-time data is mostly seasonal and non-linear.(Ly et al., 2022) has introduced the potential application of Seasonal Autoregressive Integrated Moving Average (SARIMAX), to predict wastewater quality. Ten different wastewater parameters data was collected hourly over a year from three different WWTP in China to predict outlet TP. Irrespective of WWTPs, SARIMAX consistently demonstrated the good performance with high R^2 value as 0.93.

2.1.4. Multiple Linear Regression (MLR)

Linear regression models have similar difficulties to ARIMA, that it works only with linear data. In (Rahmat et al., 2022) MLR was performed to develop a prediction approach using wastewater quality index (WWQI) for a regional WWTP in Melaka, Malaysia. Seven principal components analyses were derived with daily data and the coefficients of the WWQI model are directly dependent on influent biological oxygen demand (BOD), effluent BOD, influent chemical oxygen demand (COD), and effluent COD values. The experimental results showed that the model performed well with R^2 as 0.85. (Nair et al., 2022) presents development of MLR (Linear Regression, Ridge Regression, Baysian Ridge and Lasso Regression) to estimate

Total Phosphorus (TP) and Chemical Oxygen Demand (COD) in the influent and effluent streams of a full-scale WWTP in Norway. Ridge algorithm shows relatively better results with R^2 =0.86 for influent TP and 0.72 for influent COD. (Wang et al., 2021) tested four linear regression models (Linear Regression, Ridge Regression, ElasticNet Regression and Lasso Regression) to predict flow rate, COD and ammonia using the historical data obtained from a WWTP located in western China. The accuracy of all the models (86, 82 and 74 % for flow rate, COD and ammonia, resp.) was not sufficient for WWTP operators.(Andreides et al., 2022)

2.1.5. SVM and Boosted tree

(Wodecka et al., 2022) presents the use of classification models such as **support vector machines** and **boosted trees** methods to predict the variability of wastewater quality at the inflow to wastewater treatment plants in Poland. TP and TN was predicted on based of biochemical oxygen demand, chemical oxygen demand, total suspended solids, and ammonium nitrogen and can be identified with sufficient accuracy. Besides SARIMAX, (Ly et al., 2022) presented other shallow ML based regression models such as Random Forest (RF), Support Vector Machine (SVM), Gradient

Tree Boosting (GTB) to predict outlet TP. Result with RF, SVM, GTB, and ANFIS were unable to address large datasets with nonlinear and nonstationary behavior.

2.1.6. Neural network

The most popular non-linear models for influent characteristics prediction are based on various types of artificial neural networks (ANN). The robustness of the ANN strongly depends on the type of ANN and the quality of the training dataset. The use of daily influent data can lead to a high prediction accuracy.(Andreides et al., 2022) . In comparison with ARIMA, (Zhang et al., 2019) developed an artificial neural network model (i.e., the multilayer perceptron neural network, MLPNN) for predicting wastewater inflow. The result of MLPNN was satisfied with R^2 value as 0.79 for training dataset and 0.64 for testing dataset.

A taxonomy of literature review has been presented in appendix section A in table A.1.

3. Materials and Methods

3.1. Study sites description

3.1.1. Hias IKS Wastewater Resource Recovery Facility

Hias IKS is an inter-municipal wastewater transport, treatment and resource recovery facility owner, and a water and sewage service provider for the Hamar, Løten, Ringsaker, and Stange municipalities. Hias owns and operates the water and sewage treatment plants in Stange, with approximately 230 kilometers of transmission line, 11 pumping stations, six high basins, and six measuring stations. Hias Water and sewage play an important role in nature's cycle. All water that Hias processes and distributes to the municipalities and their networks is obtained from lake Mjøsa . At the same time, all treated wastewater is returned to Mjøsa. Hias supplies water to approx. 56,000 people and cleans sewage from approx. 65,000 people. Total water production capacity of Hias WWRF is 6.13 million m^3 and sewage quantity added to the

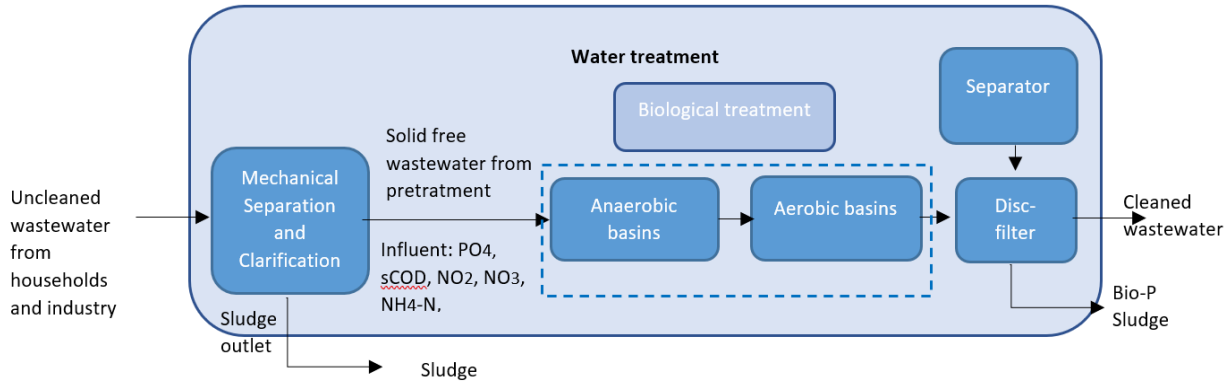


Figure 3.1: Hias wastewater treatment process block diagram

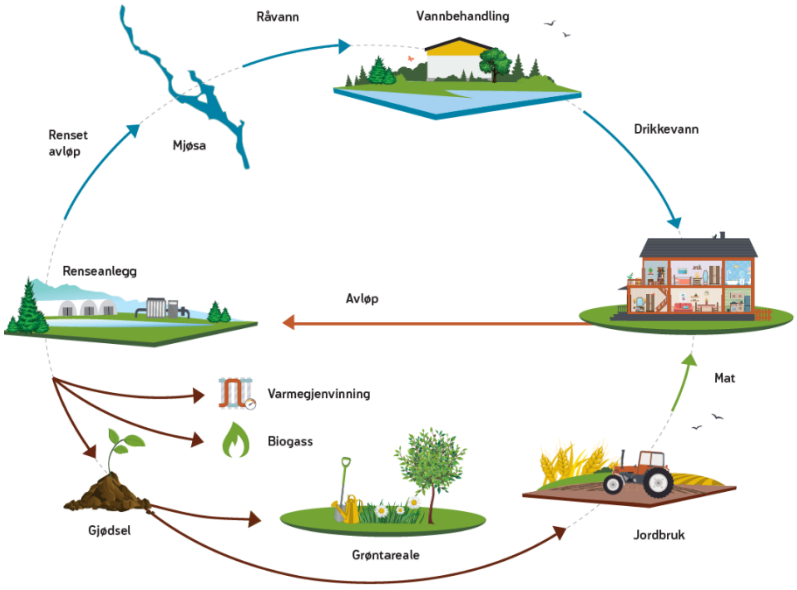


Figure 3.2: Hias wastewater resource recovery process (Photo: Hias process)

treatment plant is 6.97 million m^3. The treatment plant has a treatment capacity of 140 000 p.e. and 3 MW heating capacity from outlet. Hias process is a compact biological nutrient removal technology that removes phosphorus and nitrogen with the help of micro-organisms that grow as a biofilm on small plastic chips in the basins and circulate the basins with the wastewater. This moving bed bioreactor process with enhanced biological phosphorus removal consists of ten basins (three anaerobic basins and seven aerobic basins) and small biofilm carriers. The influent at Hias treatment plant is strongly affected by emissions from the food industry which can amount to up to 50 percent of the load. Wastewater from the food industry, such as slaughterhouses and the like, will add nutrients, organic matter, fat, proteins, blood, faeces, and washing solutions to a greater extent than typical household wastewater (https://www.hias.no/om-hias).

The Hamar city's weather varies in summer and winter, with the average temperature varying from -7° to 16° and precipitation ranges between 0.7mm to 120mm. The climate typically features two distinct seasons - High temperature and high precipitation- mainly in the summer (May to September), while the dry season lasts from October to April (https://www.yr.no).

3.1.2 VEAS process:

The VEAS wastewater treatment plant is Norway's largest treatment plant and is a crucial contributor to keep the Oslo Fjord clean. The wastewater of different municipalities, equal to that of 867,000 people, is transported through the VEAS tunnel from the discharge point to the Asker treatment plant. VEAS not only purifies water effectively but also contributes to the nutrient cycle by producing stabilized, hygienic, and lime-enhanced sewage sludge (known as VEAS soil) and a nitrogen solution beneficial for agriculture. (https://www.veas.nu/om)

VEAS processes 2,300-3,000 liters per second on a dry weather day. When it rains, more water comes to VEAS. The treatment plant can handle up to 11,000 liters per second and, in addition, store up to 200,000,000 liters in the tunnel. VEAS processes 100-110 million cubic meters of wastewater annually.

VEAS operates the waste tunnel that leads the wastewater from Fagerlia in the east to the treatment plant at Bjerkås in Asker. Oslo, Bærum, and Asker municipalities are responsible for the sewage network from the residential areas that collect wastewater that leads to the VEAS tunnel.

In addition to wastewater, this network is loaded with rainwater and snowmelt water, also called stormwater.

Water treatment

Wastewater is pumped up from the inlet pump station, located 23 meters below the treatment plant. Rags, plastic, cotton swabs, and other rubbish are removed using a sieve. The sewage waste, the screening material, is delivered to an approved landfill. Heavier particles such as sand and coffee grounds are removed from the wastewater in an aerated sand trap. To remove phosphorus and organic matter, chemicals are added, which cause small particles to bind into larger particles that sink down and form sludge. This sludge is pumped out from the bottom of the pool, while the water is carried on to biological purification, where nitrogen is removed with the help of bacteria. The biological cleaning step consists of biofilm processes with leca as carrier material. First, nitrification occurs in aerated pools, and finally, denitrification occurs with methanol, an external carbon source. The Leca material also provides good filtration of the water.

The purified water is discharged into the Oslo Fjord at a depth of 50 meters via an outlet tunnel, and five diffusers distribute the water so that it is stored at a depth of around 20 meters. The backwash water from the biological purification step is returned to the VEAS inlet.

The VEAS process has a capacity of around 3.2-3.5m3/s with normal wastewater. Up to 5.1-5.2 m3/s can be treated on increasingly diluted wastewater. In the case of diluted wastewater, the VEAS plant can process up to 11 m3/s.

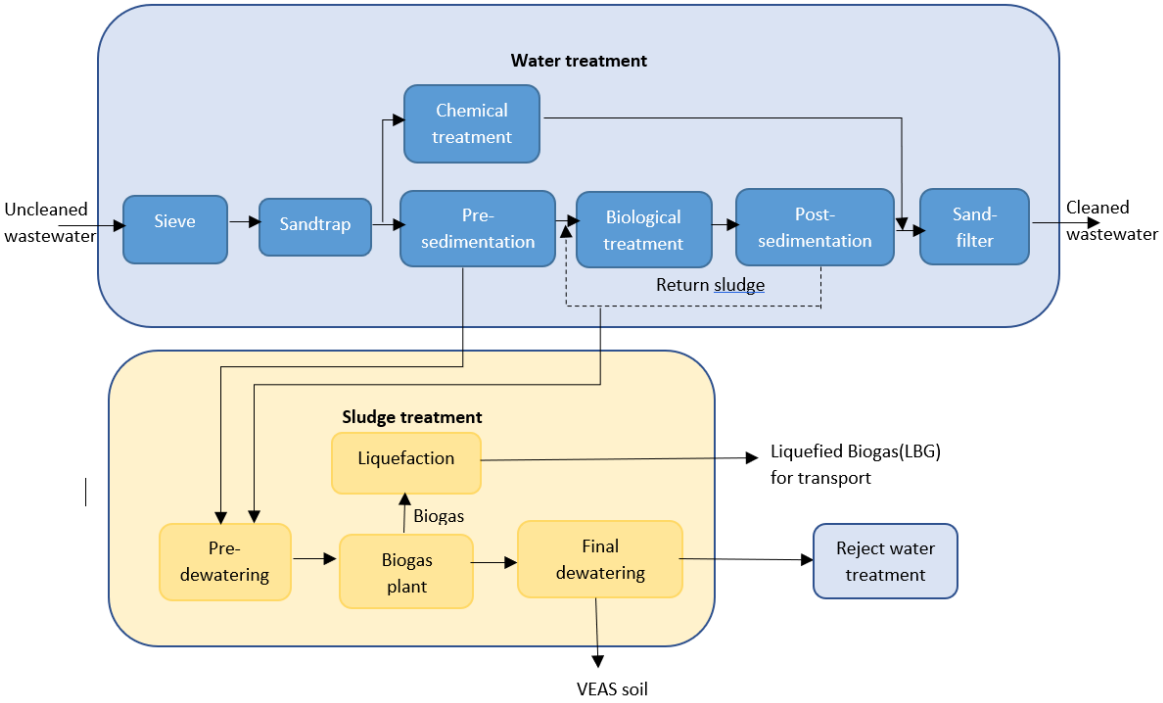


Figure 3.3: VEAS water and sludge treatment process(Jonassen et al., n.d.)

Sludge treatment

The sludge from the purification process is fed to the bioreactors. Particulate matter such as sludge is pumped from the sedimentation basins2 in the main facility via pre-dewatering to the eradication plant. Here, the sludge is broken down, and biogas is formed. The decay process produces approximately 1,400-1,500 Nm3 of biogas per hour, corresponding to an energy amount of about 70 GWh per year. Until 2020, VEAS used biogas for energy and heat in facilities and administration buildings. However, after the new refinery was built, the biogas is upgraded and became liquid fuel for the transport sector.

After the sludge has decayed, lime is added and dewatered/vacuum dried. When the sludge has been treated, it is dry and free of infectious agents and weed seeds. The end product, VEAS soil, is rich in phosphorus, lime, and organic matter and is used as a soil conditioner in grain-growing areas. Agriculture annually receives around 38,000 tons of Veas's soil.

The water squeezed out of the sludge is rich in ammonia and cleaned in the stripping and absorption plant. An ammonium nitrate solution of around 4,000 tons per year is produced, which can be used as fertilizer or industrial fertilizer raw material.(Jonassen et al., n.d.)

3.2. Data collections and pre-processing

Data preparation was conducted with two stages: (1) Data collection and (2) Data pre-processing to gather and transform raw data into a time series dataset for statistical and machine learning modelling and forecasting wastewater influent composition. The procedures are described in Fig. 3.4.

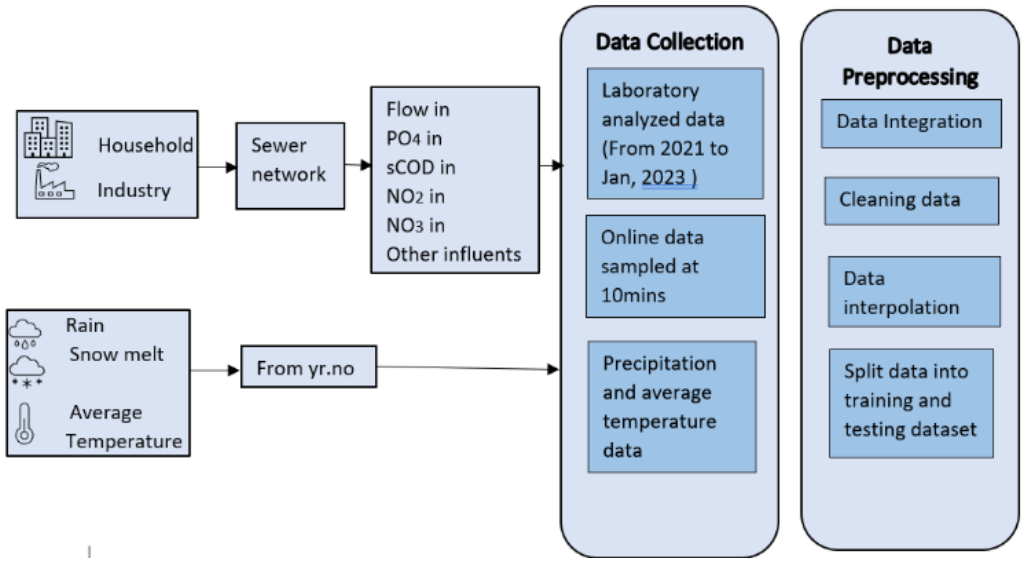


Figure 3.4: Data collection and data pre-processing procedure

The same process was applied to collect and transform raw data to usable data for sludge influent.

3.2.1 Data Collection

The data collection has been done in two phases for online and laboratory-analyzed data.

HIAS online dataset: The treatment plant is equipped with a state-of-the-art online monitoring system with remote data access capabilities**.** sCOD(mg/l), NO_2(mg/l), NO_3(mg/l) and flow-rate(l/s**)** sensors that relay real-time information to a data acquisition system. Data is collected from 1st December 2022 to 31st January 2023 with a sampling frequency 10min. Air temperature(°C) and precipitation(mm) data has been collected from The Norwegian Meteorological Institute website (yr.no) of same time with 10 min. sampling frequency. The statistical descriptive summary of the wastewater influent parameters of Hias online dataset is tabulated in Table 3.1.

Table 3.1: Statistical description of Hias online dataset (1st December 2022-31st January 2023) along with data from yr.no

	Flow rate (l/s)	sCOD (mg/l)	NO_2 (mg/l)	NO_3 (mg/l)	Temp.(°C)	Percip. (mm)
count	8752	8843	4601	7582	8852	8852
mean	89.35	460.35	0.62	2.80	-5.63	0.05
std	22.27	95.96	0.52	1.02	5.64	0.40
min	20.37	69.25	0	0.31	-20.4	0
25%	72.91	393.17	0.20	1.94	-9.73	0
50%	92.13	469.84	0.50	2.75	-5.14	0
75%	108.01	523.91	0.93	3.58	-0.66	0
max	123.60	751.76	3.39	5	6.4	10.22

HIAS Laboratory-analyzed dataset: Two years' historical data of water temperature at inlet, influent flow rate(m^3/h), inlet sCOD(mg/l), PO_4-P in(mg/l) and ammonia (NH_4-N) (mg/l) is collected from laboratory-analysed data. Parameters are analysed once a day on every weekday. The statistical descriptive summary of the wastewater influent parameters is tabulated in Table 3.2.

Table 3.2: Statistical description of Hias lab dataset (2021-2023)

	Q (m3/h)	Temperature	PO_4 in (mg/l)	SCOD in mg/l)	NH_4 in (mg/l)
count	702.0	702.0	702.0	702.0	702.0
mean	344.9	10.86	5.2	436.1	60.8
std	50.4	2.55	1.3	121.1	12.4
min	163.2	6.30	1.5	122.0	25.0
25%	310.8	8.40	4.4	350.7	54.0
50%	336.4	10.80	5.2	442.5	61.0
75%	376.7	13.30	6.2	527.0	68.4
max	502.3	15.20	9.2	804.0	120.0

VEAS Online dataset: Sludge treatment process is complex, and it is difficult to apply direct sensors at the inlet of sludge treatment process. As we are interested to predict the influent compositions of sludge treatment process it was not possible to get online sensor data for inlet composition of sludge from treatment process, which leads us to work with only laboratory analyzed data of sludge influent composition.

VEAS Laboratory-analyzed dataset: One month data (30th June to 29th July 2022) are collected from lab analyzed data from the inlet of pre-dewatering section. Dataset consists of several parameters such as Flow rate(l/s), sCOD(mg/l), NH_4-N(mg/l), pH, TS%(Total solid), Organic acid(meq/l), total alkalinity(meq/l), Protein, raw fat, carbohydrate etc. These parameters were analyzed at VEAS laboratory once a day on every weekday. Statistical description of the monitored sludge influent parameters at VEAS is shown in Table 3.3.

Table 3.3: Statistical description of VEAS lab dataset(June-July, 2022)

	Flow rate(l/s)	SCOD (mg/l)	NH_4-N (mg/l)	pH	TS%	Organic acid (meq/l)	Total alkalinity	Protein	Raw fat	carbohydrate
count	673	673	673	673	673	673	673	673	673	673
mean	6.78	9438.61	645.56	5.94	6.81	101.43	75.50	19.21	0.88	24.17
std	1.16	1608.30	198.29	0.38	0.92	19.92	15.06	2.48	0.13	3.68
min	4.59	6575.00	374.20	5.33	4.45	70.90	53.61	12.69	0.53	15.08
25%	5.99	8207.23	512.32	5.73	6.20	85.97	67.58	17.78	0.78	22.55
50%	6.90	9348.48	592.47	5.85	6.76	99.45	72.76	19.03	0.88	24.16
75%	7.54	10567.0	721.15	6.04	7.65	110.90	79.33	20.66	0.99	25.73
max	9.380	14018.0	1395.20	7.38	8.25	160.7	132.0	24.45	1.27	36.84

3.2.2 Data pre-processing

Data integration combines data from multiple sources into a single dataset for analysis. The online monitored dataset of Hias was combined with precipitation and average temperature data from yr.no. Parameters chosen from the laboratory dataset are Flow rate at inlet Q (m3/h), Water temperature (°C), PO_4 in(mg/l), sCOD in(mg/l) and NH_4(mg/l). In addition, we have combined precipitation and average temperature data from yr.no with laboratory data.

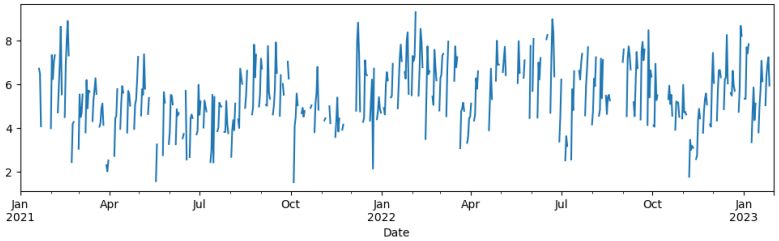
Data transformation involves converting the data into a suitable format for analysis. This step may include normalization, standardization, encoding categorical variables or resample into suitable time frequency. For example, VEAS lab analysed data of daily frequency is resampled to time-series data with sampling frequency of one-hour to make better prediction model with SARIMAX. It increases the size of dataset from 30 rows to 697 rows.

df_hourly = raw_data.resample('H').interpolate()

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 673 entries, 2022-07-01 00:00:00 to 2022-07-29 00:00:00
Freq: H
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Flow rate(l/s)     673 non-null    float64
 1   sCOD FOR(mg/l)     673 non-null    float64
 2   NH4-N FOR          673 non-null    float64
 3   pH FOR             673 non-null    float64
 4   TS% FOR            673 non-null    float64
 5   Organiske syrer    673 non-null    float64
 6   Total alkalitet    673 non-null    float64
 7   Protein FOR        673 non-null    float64
 8   Råfett FOR         673 non-null    float64
 9   Karbohydrater FOR  673 non-null    float64
```

Data cleaning is conducted to identify and correct or remove probable inaccurate or irrelevant data. Many data errors are detected, including non-numerical, abnormally large and unexpected values. These data points are called outliers. They are determined by sorting the dataset ascending and descending order. All of them are handled to achieve a more consistent and better accuracy dataset to build a predictive model for wastewater inflow and PO_4 at inlet. These outliers can be removed by python script library numpy. It can identify the outliers as data points greater than the threshold value, where **threshold** is the number of standard deviations from the mean. The filtered Hias lab-dataset with 3790 data points remaining after error elimination is converted to a daily time series dataset.

This converted dataset is then inspected to find out any missing data. As laboratory-analyzed data are only recorded once in 24 hours and from Monday to Friday, almost 1537 missing values are detected. These missing values are handled by the "spline interpolation" method. Spline interpolation is a mathematical method commonly used to construct new data points within the boundaries of a set of known points, and there was no impact on the nonlinear datasets. The remaining NaN values are handled by drop functions and at the end we get 702 days of dataset.



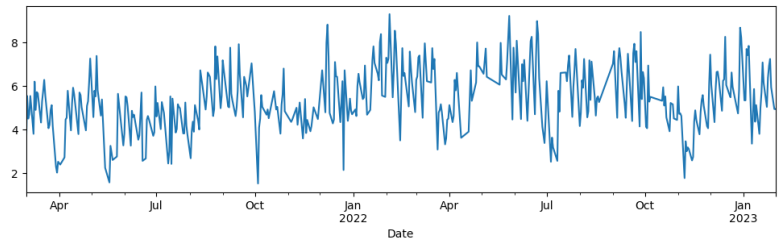


Figure 3.5: PO_4 data vs time plot (top): before interpolation, (bottom): after interpolation

From Hias online dataset the influent parameter NO_2 need to be discarded as it has 4261 missing or NaN values.

```
print(raw_data.isnull().sum())
```

```
Flow rate (l/s)     110
sCOD(mg/l)           19
No2(mg/l)          4261
No3(mg/l)          1280
temp                 10
precip               10
dtype: int64
```

After data interpolation and removing outliers and missing parameters from Hias online dataset, 8819 count of flow rate, sCOD and NO_3 parameters are further used for model development.

```
cleaned_data.describe()
```

	Flow rate (l/s)	sCOD(mg/l)	No3(mg/l)
count	8819.000000	8819.000000	8819.000000
mean	89.407816	461.694850	3.087207
std	22.163590	93.569416	1.194537
min	29.676300	175.524000	0.315300
25%	73.105150	394.569000	2.128700
50%	92.338300	470.897000	2.998000
75%	107.866800	524.051500	4.049250
max	123.604000	719.720000	5.000000

Data partition: The laboratory-analysed and online dataset of wastewater and sludge influent parameters are then divided into output, input, training, and testing datasets. As stated by Hyndman and Athanasopoulos (2018), typically, the size of the testing set accounts for around 20% of the entire dataset and is ideally at least equal to the longest forecasting duration. Therefore, the ratio of training to testing set is 80:20. The training set is used for model development, and the testing set is reserved for model validation.(Do et al., 2022)

3.3 Software Packages

The open-source programming language Python (www.python.org) was used to process the raw lab-analyzed and online sensor data, generate mathematical models, and deploy algorithms for real-time estimation. The open-source library, pandas, version 1.3 (https://pandas.pydata.org/), was used to clean and preprocess raw data and generate concurrent datasets of the same timestamp.

statsmodels is an open-source Python module (https://www.statsmodels.org/)that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. This module was used to train ARIMA and SARIMAX model with different datasets.

Scikit-learn, a free Python library (https://scikit-learn.org/) provides several algorithms to train MLR models and obtain the regression coefficients. (Nair et al., 2022)

TensorFlow is an end-to-end open-source platform for machine learning (https://www.tensorflow.org/). TensorFlow is a rich system for managing all aspects of a machine learning system. Keras (https://keras.io/) is a high-level neural network library that runs on top of TensorFlow and it was used to build and optimize the neural network model.

3.4. Data analysis

Analysis of Influent Nutrient Composition and Quantity

The first and most important requirement for the development of time-series mathematical or machine learning model is to analyse the wastewater influent parameters data. The main criterion for developing time-series mathematical models such as ARIMA, SARIMA or SARIMAX model is that time-series data is stationary. A time series is stationary when its statistical features (e.g., mean and variance) are constant over time, or not impacted by the time at which the series is observed. The term “stationarity” is used to imply the stationary status of a time series. In contrast, when a time series exhibits trends (e.g., upward or downward) or seasonal patterns (e.g., quarterly, monthly, or weekly), it is non-stationary. (R. Wang et al., 2021).

Therefore, before employing model development the stationarity of the original training time series of raw wastewater and sludge inflow data was investigated. The **ADF and KPSS tests**, and the **ACF and PACF plots** were used to verify the data’s stationarity.(Do et al., 2022)

- ADF test

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical tests with a hypothesis testing involved with a null and alternate hypothesis and as a result a test statistic is computed and p-values get reported.

The ADF test belongs to a category of tests called 'Unit Root Test', where Unit root is a characteristic of a time series that makes it non-stationary. A unit root is said to exist in a time series of the value of α = 1 in the equation: $Y_t = \alpha Y_{t-1} + \beta X_e + \varepsilon$ (3.1)

where, Y*t is the value of the time series at time 't' and X*e is an exogenous variable. the number of unit roots contained in the series corresponds to the number of differencing operations required to make the series stationary.

A Dickey-Fuller test is a unit root test that tests the null hypothesis that α=1 in the following model equation. alpha is the coefficient of the first lag on Y.

Null Hypothesis (H0): α=1

$y_t = c + \beta t + \alpha y_{t-1} + \phi \delta Y_{t-1} + e_t$ (3.2)

where, y(t-1) = lag 1 of time series

delta Y(t-1) = first difference of the series at time (t-1)

the coefficient of Y(t-1) is 1, implying the presence of a unit root. If not rejected, the series is taken to be non-stationary.

The ADF (Augmented Dicky-Fuller) test expands the Dickey-Fuller test equation to include high order regressive process in the model, can be expressed as:

$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \delta Y_{t-1} + \phi_2 \delta Y_{t-2} + \cdots + \phi_p \delta Y_{p-1} + e_t$ (3.3)

Since the null hypothesis assumes the presence of unit root, that is α=1, the p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis. Therefore, inferring that the series is stationary.(Selva Prabhakaran, 2019)

The statsmodel package in python provides a reliable implementation of the ADF test with from statsmodels.tsa.stattools import adfuller . It returns:

1. The p-value
2. The value of the test statistic

3. Number of lags considered for the test.
4. The critical value cutoffs.

- KPSS test

Like ADF test, the KPSS test is also commonly used to analyse the stationarity of a series. However, it has couple of key differences compared to the ADF test in function and in practical usage. The KPSS test, short for, Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a type of Unit root test that tests for the stationarity of a given series around a deterministic trend. A key difference from ADF test is the null hypothesis of the KPSS test is that the series is stationary. So practically, the interpretation of p-value is just the opposite to each other.

That is, if p-value is < significant level (say 0.05), then the series is non-stationary. In python, the statsmodel package provides a convenient implementation of the KPSS test.(Selva Prabhakaran, 2019)

```
from statsmodels.tsa.stattools import kpss
```

- Differencing

After checking the stationarity of the training time series using statistical tests, the non-seasonal differencing *d* and seasonal differencing *D* were determined. If the series is stationary, it is not required to execute the process of differencing, and the value of parameters *d* and *D* is zero. In case the series is non-stationary with the presence of seasonality and trend, the seasonal difference is applied. When there is no trend and seasonality component, the series is transformed by the non-seasonal difference. The value of parameters *d* and *D* implies the number of times the wastewater inflow series needs to be different to satisfy stationarity. (Do et al., 2022)

- ACF/PACF

The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the original training time series are created if required to further confirm its stationarity.

In this study, the ACF plots depict the correlation coefficient between the wastewater influent parameter (ex. PO_4 and sCOD for Hias dataset) and sludge influent parameter (ex. Organic acid for VEAS dataset) time series and its own lagged values, and the PACF plots measure the partial correlation coefficient between this data series and lagged versions of itself. The next step was to

plot the ACF and PACF of the stationary time series. It could be the original training time series with stationary status or the differenced series after differencing process obtained from the previous step. The non-seasonal and seasonal orders of AR (parameters p and P) and MA (parameters q and Q) were identified based on the ACF and PACF plots. Different values of those parameters were combined to identify possible configurations of (p,d,q) and (P,D,Q) for potential ARIMA or SARIMAX models.(Do et al., 2022)

- AIC/BIC

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are statistical metrics used to evaluate the goodness-of-fit of an ARIMA or SARIMA model. Both AIC and BIC provide a measure of the quality of a model while penalizing for model complexity.

AIC is defined as follows:

$AIC = 2k - 2\ln(L)$ (3.4)

Where, k is the number of parameters in the model, and ln(L) is the log-likelihood of the data given the model. AIC is a measure of the trade-off between the goodness-of-fit of the model and the number of parameters used in the model. The lower the AIC, the better the model fits the data.

BIC is similar to AIC but places a more substantial penalty on models with more parameters. BIC is defined as:

$BIC = k\ln(n) - 2\ln(L)$ (3.5)

Where n is the number of observations in the data. BIC measures the relative quality of the model compared to other models under consideration. The lower the BIC, the better the model fits the data.

AIC and BIC can be used to compare different ARIMA or SARIMA models and choose the best one for a particular dataset. The model with the lower AIC or BIC value is generally preferred when comparing models. (https://fr.mathworks.com/)

3.5. Model Development

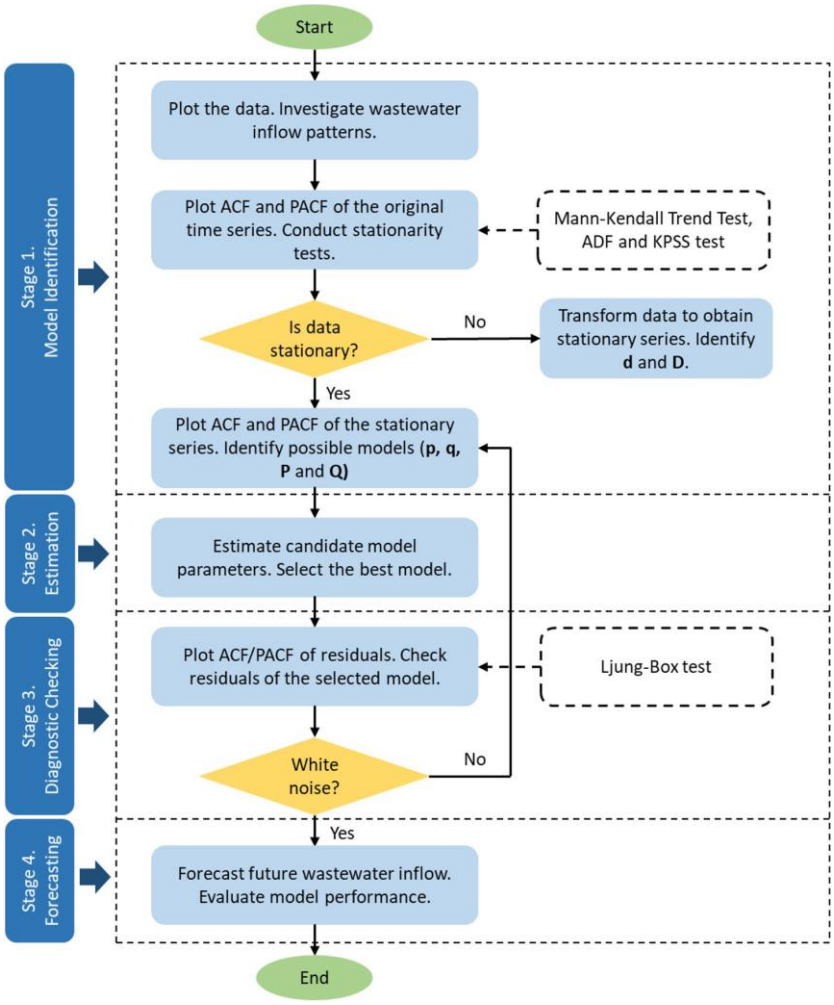


Figure 3.6: Flow chart of general model development process for ARIMA and SARIMAX

Figure. 3.6 illustrates the flowchart of the step-by-step methodology applied for modelling and forecasting wastewater inflow COD and PO_4 concentration at inlet. The procedures are based on Box and Jenkins methodology. (Do et al., 2022)

Stage 1. Model identification

After data pre-processing and data analysis, the first step of the Box-Jenkins model development process is the identification stage.

a) For visualization, we need to plot and decompose the time series into its components to observe any trends, seasonality, and residuals.

b) The autocorrelation (AC) and partial autocorrelation (PAC) graphs are usually used to determine if there were any signs of trends within the series. When no trends are present in the series, the AC and PAC graphs quickly converge to zero, meaning that value of (d), which denotes the number of times that the observation data are differenced, can be determined for the ARIMA model.
c) Along with AC and PAC plot, Augmented Dickey-Fuller (ADF) test or Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test has been done to check if the time series is stationary or non-stationary.
d) If the time series is non-stationary, apply differencing to make it stationary. Determine the order of differencing (d) required by finding the minimum value of d that makes the time series stationary.

Stage 2. Parameter estimation

In this stage, various potential models identified in stage 1 were examined. If the time series exhibits seasonal patterns, then a seasonal ARIMA (SARIMA) model is more appropriate. Determine the seasonal period (s) and repeat previous steps to identify the seasonal AR and MA terms. In this way, determine the order of the seasonal component (P, D, Q) required to make the time series stationary. Use the seasonal decomposition plot to help determine these parameters. The coefficient of determination (R^2), root mean square error (RMSE), and normalized Bayesian information criterion (BIC) were used to select one amongst the potential models. The best model with the optimal set of parameters has the highest R^2, and the least RMSE and normalized BIC and lowest AIC value. (Do et al., 2022)

Fit the model: Use the identified orders of p, d, and q to fit the ARIMA or SARIMA model to the data. This has been done using python software packages **statsmodels** library.

Stage 3. Diagnostic checking

Diagnostic tests such as residual plots, ACF and PACF plots, and Ljung-Box tests are used to evaluate the model's goodness-of-fit. If the model does not fit the data well, try adjusting the

orders of p, d, and q and repeat stage 2(parameter estimation) until a satisfactory model is obtained.

3.4.1 ARIMA model

For **ARIMA** model, for the training set, three parameters, including p, d, and q, were configured manually. The integrated value d was first found in the identification process, then a number of (p, q) sets were searched using a grid search algorithm. For each (p, q) set, the other coefficients in the ARIMA model were estimated with fixed p, q, d values. The optimal (p, d, q) combination would be found by choosing the set with the lowest root mean square error (RMSE) and lowest AIC value. Searching for the optimal combination of (p, d, q) could help calibrate the model for best performance. Once the best combination of (p, d, q) was found and the other coefficients were calibrated, the model could be finalized and loaded to make one-step ahead predictions for the testing period. (Boyd et al., 2019)

Stage 4. Forecasting

Once a satisfactory model has been obtained, use it to make predictions for future time periods. A model with the highest accuracy in simulating wastewater composition PO_4 at inlet would be employed to forecast data. Applying the selected ARIMA model, the wastewater time series are forecasted using the python software function **model.predict**(data). The predicted values are then matched against the testing set and evaluate model performance in terms of R^2 and RMSE.

3.4.2. SARIMA and SARIMAX:

SARIMA is developed by including additional seasonal components to the ARIMA model, which handles the seasonality in the time series. SARIMA model, in general, is a combination of the non-seasonal module (p,d,q) and seasonal module (P, D, Q)s with seven parameters. It is denoted as SARIMA (p, d, q)(P, D, Q)s ; where p and P is the order of non-seasonal and seasonal AR(Auto Regression) terms; d and D is the degree of non-seasonal and seasonal differencing; q and Q are the order of non-seasonal and seasonal MA (Moving Average) term and s is the length of seasonality in the time series. For example, in an hourly time series, s = 24; in a daily time series, s = 7; in a monthly time series, s = 12; and in a quarterly time series, s = 4.(Do et al., 2022)

SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with Exogenous factors) is an updated version of the SARIMA model. SARIMAX is a seasonal equivalent model like SARIMA and Auto ARIMA, which can also deal with external effects. This feature of the model differs from other models.

3.4.3. Multiple linear regression (MLR):

As discussed in the theoretical background section multiple regression models have been developed with different datasets and evaluated their accuracy. Regression models are used to describe relationships between variables by fitting a line to the observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for n explanatory variables x1, x2, ... , xp is defined to be

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon. \quad (3.6)$$

where y = the predicted value of the dependent variable, β0= the y-intercept (value of y when all other parameters are set to 0), β1x1 = the regression coefficient (β1) of the first independent variable (x1) (the effect that increasing the value of the independent variable has on the predicted y value).

βnxn = the regression coefficient of the nth independent variable. = model error (how much variation there is in estimation of y) (Nair et al., 2022)

For purpose of this study, y represents PO_4 in Hias lab dataset, sCOD in Hias online dataset and Organic Acid in VEAS lab dataset. x1, x2 …xn represents sCOD, Flow rate and NH_4-N in Hias lab dataset, Flow rate and NO_3 in Hias online dataset and flow rate, NH_4-N, pH and TS% in VEAS lab dataset.

To increase prediction accuracy, high order terms and interaction terms can be involved as inputs. In this study, square terms of original variables and two-effect-interaction terms were applied in MLR models. All the original variables, square terms and interaction terms were included to train an over-fitted model at the first step. (X. Wang et al., 2019)

In this study, we have added the previous output steps and all input steps to improve the accuracy of regression models.

Including the previous output step as a feature can capture temporal dependencies in the data. As in real-life time-series data, where data are collected over time, the previous step's output can

provide useful information about the current observation, which can help the model better capture trends and patterns in the data, resulting in more accurate predictions.

Including all inputs ensures that the model captures all relevant information about the problem. Omitting important features can lead to underfitting, where the model needs to be more complex and capture the complexity of the data. By including all inputs, the model has a better chance of accurately modeling the relationships between the inputs and the output. Also, including all inputs in the regression model can reduce the impact of confounding variables. Confounding variables correlate with both the inputs and the output but are not themselves inputs. By incorporating all inputs, the model can effectively separate the impact of individual variables on the output, leading to more precise predictions.

However, it is important to note that adding too many features can also lead to overfitting, where the model becomes too complex and performs poorly on new, unseen data.

3.4.4. Neural Network

The ANN (Artificial neural network) is fed by a series of input-output pairs and is trained to reproduce the outputs. The learning process of a feedforward network occurs in the perceptron, which consists of a single neuron with adjustable weights and an activation function. When multi-layer feedforward network has perceptron with two or more trainable weight layers, it might also be called an MLPNN (Multilayer perceptron neural network).(Zhang et al., 2019)

In this study, an MLPNN model was developed for wastewater and sludge influent composition forecasting. First, in Hias laboratory-analyzed dataset, The inputs were prior daily flow rate data, daily inlet PO_4 concentration, daily inlet sCOD concentration and daily inlet NH_4-N.

In Hias online dataset, the inputs were prior flow rate data, inlet sCOD concentration and inlet NO_3 concentration data with a sampling frequency of one hour.

In VEAS laboratory-analyzed dataset, the inputs were prior hourly flow rate data, hourly inlet organic acid concentration, hourly inlet NH_4-N concentration , hourly inlet pH measure and hourly inlet TS(Total suspended solid) percent data.

A four-layer MLPNN was selected as the best model structure, based on a series of exploratory experiments. The exploratory experiments manually searched and tested different numbers of hidden layers, neurons, and epochs.

In this study, Adam optimizer has been used, with a learning rate of 0.01. Adam (short for Adaptive Moment Estimation) is an optimization algorithm that is commonly used for training neural networks. Like other optimization algorithms, Adam works by iteratively adjusting the weights of a neural network to minimize the loss function during training. Adam uses a moving average of the gradient and the second moment of the gradient to adaptively adjust the learning rate for each weight in the neural network. This allows it to converge faster and more reliably than traditional stochastic gradient descent methods. Additionally, Adam can handle sparse gradients, which makes it well-suited for problems with large datasets or sparse features.

The MLPNN used in this study consists of one input layer with the same shape as the number of features in the training data, followed by two hidden layers with 15 and 10 nodes respectively, both using the ReLU activation function. The output layer has a single node and uses the linear activation function. The loss function used for training is squared error (MSE). The epoch used in this study is 100.(Zhang et al., 2019)

ReLU activation function returns 0 for negative inputs and the input itself for positive inputs. Mathematically, ReLU is defined as f(x) = max(0,x). This function can help the model learn faster due to its non-linear nature.

Linear activation function, on the other hand, is a simple activation function that returns the input as is. It is defined as f(x) = x. This function was used in the output layer because it allows the model to predict continuous values without any range limitations.

After initializing input, output, and hidden layers a python function was called that performs k-fold cross-validation on the neural network model.

- The function first takes the training data **X_train** and **y_train**, the number of folds **num_folds**, the number of epochs **n_epochs** for training the model, and a model instance **model** as input.
- The function initializes a K-Fold object with the specified number of folds, and then iterates over each fold to train and evaluate the model.

- For each fold, the function compiles a new neural network model instance using the **compile_NN** function, and then creates a **Pipeline** object that first scales the input data using **StandardScaler()** and then applies the neural network model.
- The model is trained using the training data for the current fold, and then evaluated using the validation data. The evaluation metrics used are the R-squared coefficient (**R^2**), root mean squared error (**rmse**), and mean absolute error (**mae**).

This method provides a way to evaluate the performance of a neural network model using k-fold cross-validation, which can help to assess the model's generalization ability and prevent overfitting.

3.6. Model performance evaluation

There are several error measures that could be used for making comparisons between observed and predicted time series. Since no one measure is superior on all criteria, different. measures were used for model evaluation. To determine the accuracy of the different models in wastewater and sludge influent predictions, the mean-square error (MSE), mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE) and the coefficient of determination (R^2) are used as statistical indicators to evaluate the fit of the forecasted to the observed values.

RMSE is the square root of the average of square differences between the actual and predicted values at specific timestamps. RMSE describes the magnitude of the error which could be useful to decide the accuracy of a forecasting method. Can be described as :

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{a}^{N}(X_a - Y_a)^2} \qquad (3.7)$$

where X_a is the actual value, Y_a is the predicted value, and N is the total number of data points. RMSE depends on a few different factors. RMSE depends on the units and the frequency. of the dataset, meaning it cannot be absolutely defined as a good or bad value. If the RMSE is the same as the standard deviation, the model would only be as accurate as using the mean as the prediction. Hence, if the RMSE value is lower than the standard deviation, it implies.

that the model predicts the data better compared to using the mean as the prediction. RMSE also depends on the variance between the actual and predicted values because the difference is squared. Therefore, the primary benefit of RMSE is that it gives high weights to larger deviations, which in return, represents a better model performance. However, a disadvantage includes the fact that outstanding outliers can heavily skew the RMSE results and show a misleading model performance. (Boyd et al., 2019)

MAPE has also been used widely in scientific research as it is simple to use. MAPE is the mean of the individual theoretical errors calculated at each timestamp, as seen in Equation.

$$\text{MAPE} = \frac{1}{N}\sum_{a}^{N}\frac{|X_a - Y_a|}{X_a} * 100 \qquad (3.8)$$

The lower values of RMSE and MAPE imply a more reliable and robust model.

The coefficient of determination (R-squared) is another error criterion which was used for model validation. R-squared has been used frequently for model evaluation. It is a statistical measure which represents the ability of the independent variable (observed) to predict the variations of the dependent variable (predictions). Therefore, the correlation is between the line of best fit and the predicted values. The closer the R-squared is to the value of one, the better the model has performed as there is less error variance. The opposite is said to be as the R-squared approaches zero. In general, an R-squared value greater than 0.5 is acknowledged as being acceptable. (Boyd et al., 2019)R-squared is the sum of the distance between the predicted value and the linear line, divided by the sum of the distance between the predicted value and the mean of predictions and can be expressed as:

$$R^2 = 1 - \frac{\sum_{a}^{N}(X_a - Y_a)^2}{\sum_{a}^{N}(\bar{X} - \bar{Y})^2} \qquad (3.9)$$

where *N* is the total number of measurements, *Xa* are the actual values, *Ya* are the corresponding estimated values, $\bar{X}$ is the mean of the actual values of the X variables, and $\bar{Y}$ is the mean of the estimated values in Y variables.(El-Rawy et al., 2021)

Python library scikit learn is used to find these parameters: **sklearn.metrics**.r2_score

4. Results

Real time data of sCOD, Flow rate and NO_3 from online measurement (with additional data of average temperature and precipitation from yr.no) and lab- analyzed values of PO_4, sCOD, flowrate and NH_4-N were obtained from Hias process system. Laboratory- analyzed data of Flow rate(l/s), sCOD(mg/l), NH_4-N(mg/l), pH, TS%(Total solid), Organic acid(meq/l), total alkalinity(meq/l), Protein, raw fat, carbohydrate has been used from VEAS process. Result section is divided mainly into the three sections:

- Result for HIAS laboratory analysed data
- Result for HIAS online data
- Result for VEAS laboratory analysed data

Each section is divided into several subsections to show result of analyzed data, result of model developing stages, best model to fit for time series analysis such as ARIMA or SARIMA, Prediction plot with ARIMA or SARIMA, Predicted vs. Actual plot for different Regression methods and Neural network method. Lastly all model performance evaluation has been tabulated in measure of R^2 and RMSE.

4.1. Result HIAS laboratory analysed data

4.1.1. Result of data analysis

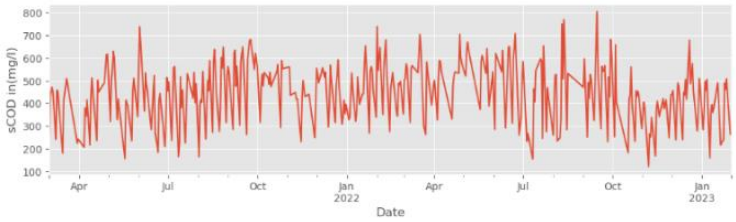


Figure 4.1: The influent characterstics of Hias WRRF during 2021-2023 (a)Flow rate, (b)PO_4 in, (c) NH_4 in, (d) sCOD in

As shown in Figs 4.1 a-d, there was no drastic change in influent quantity during 2021-2023, and the average influent flow from May to August (wet season) was higher than that from November to March (dry season), and the highest one-day inflow occurred in May 2021. Some higher values are found in the dry winter season, which could be caused by snow melting. The average influent PO_4 value in 2021 was lower than in 2022. Influent quality was typically higher from November to March (dry season) and lowered between May and August (wet season) because of much-diluted nutrients in the rainy season. The highest one-day value occurred in January 2022, and the lowest was in October 2021. From 2021 to 2023, the average influent COD value was stable. There were no significant higher or lower values, and it did not show any seasonality. The average values of influent NH_4-N were stable from 2021 to 2023, while the maximum increased in 2022. Compared with 2021, the maximum NH_4 concentration in August 2022 was much higher. Overall, the influent flow rate and other parameters concentration data are non-linear in nature. Inflow rate and PO_4 concentration data have shown a certain seasonality, however we can see the data are moreover stationary in nature.

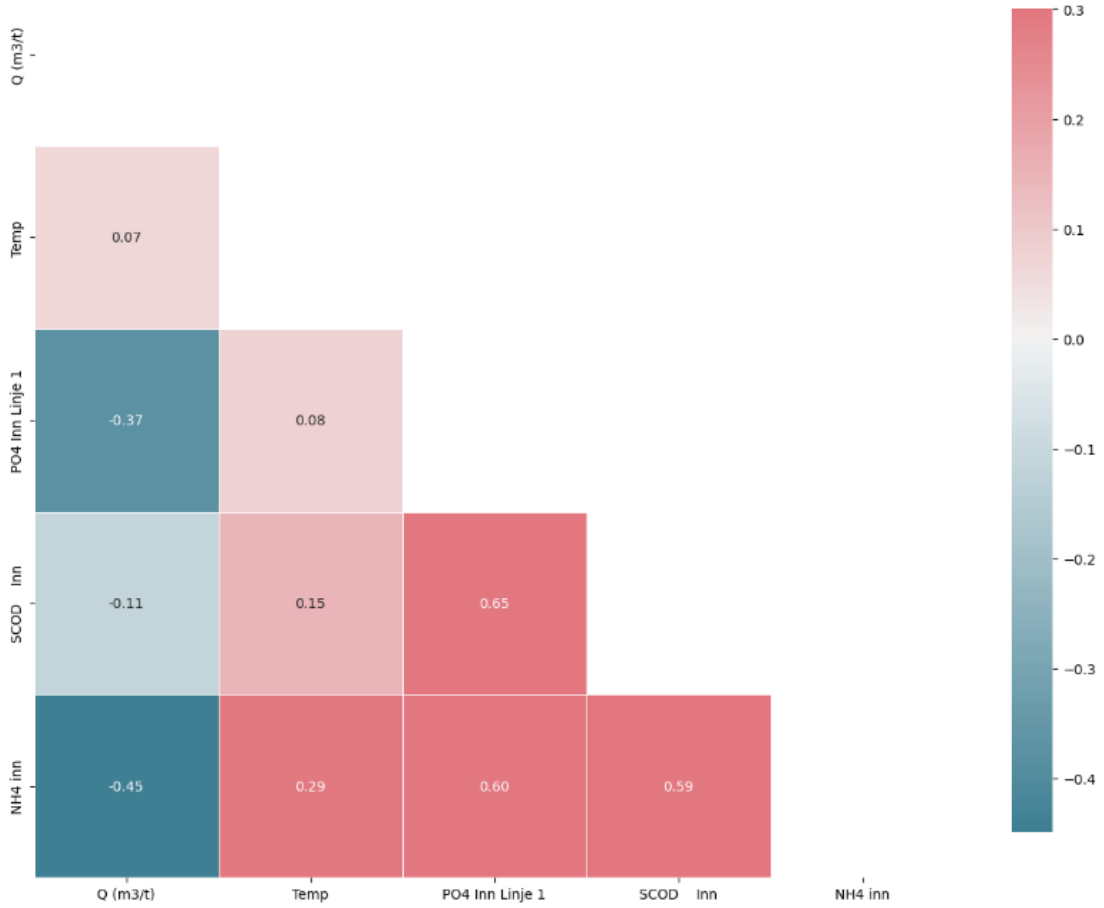


Figure 4.2: Correlation matrix and heat map between different parameters from Hias lab dataset

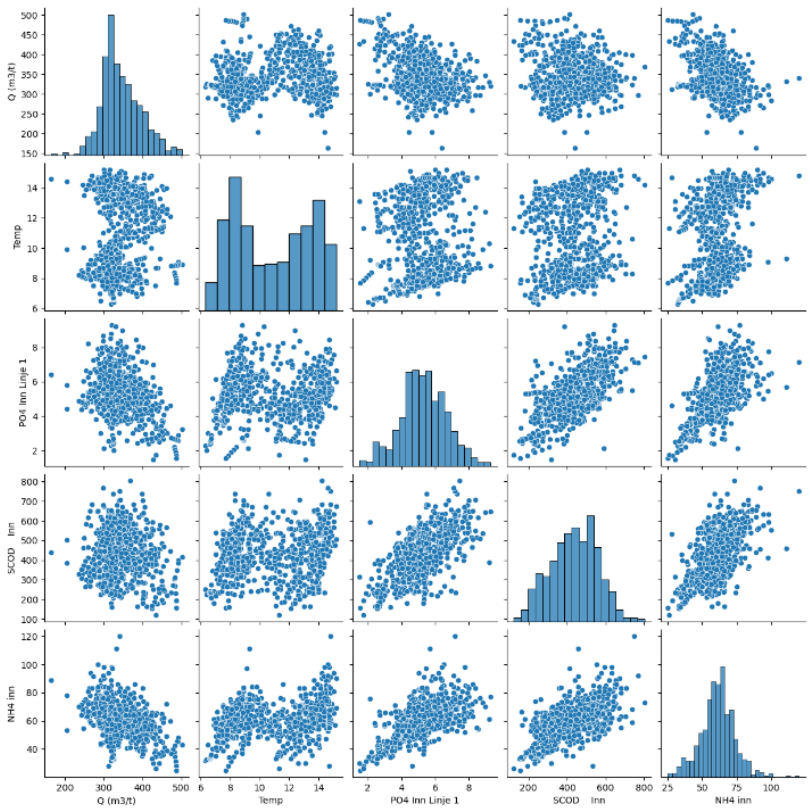


Figure 4.3: Pairplot between different parameters from Hias lab dataset

The correlations among wastewater-influent parameters collected for Hias laboratory-analyzed data is illustrated in Fig. 4.2. It shows that flowrate is negatively correlated with inlet composition PO_4, sCOD, and NH_4 means more diluted wastewater has less concentration of influent composition. The temperature parameter was excluded further for the model building of the poor correlation between inlet water temperature and compositions. The positive relationship of inlet PO_4, NH_4-N and sCOD is confirmed with correlation coefficient greater than 0.5, meaning that influent sCOD concentration increased proportionally to inlet PO_4. Thus for inlet PO_4 prediction flow rate, sCOD and NH_4 are used as input parameters whereas for inlet sCOD prediction PO_4, NH_4, and flow rate are used as input parameters.

4.1.2. Result of Model Development

Stage 1. Model identification

Before employing the ARIMA technique to develop a forecasting model, the time series data needs to be in a stationary condition. Therefore, the stationarity of the original training time series of daily inlet PO_4 concentration of WWRF was investigated.

- Plotting and decomposing the time series into its components to observe any trends, seasonality, and residuals with the help of seasonal_decompose function of python.

decompose_data = seasonal_decompose(df_daily['PO_4 Inn Linje 1'], model="additive", period=7)

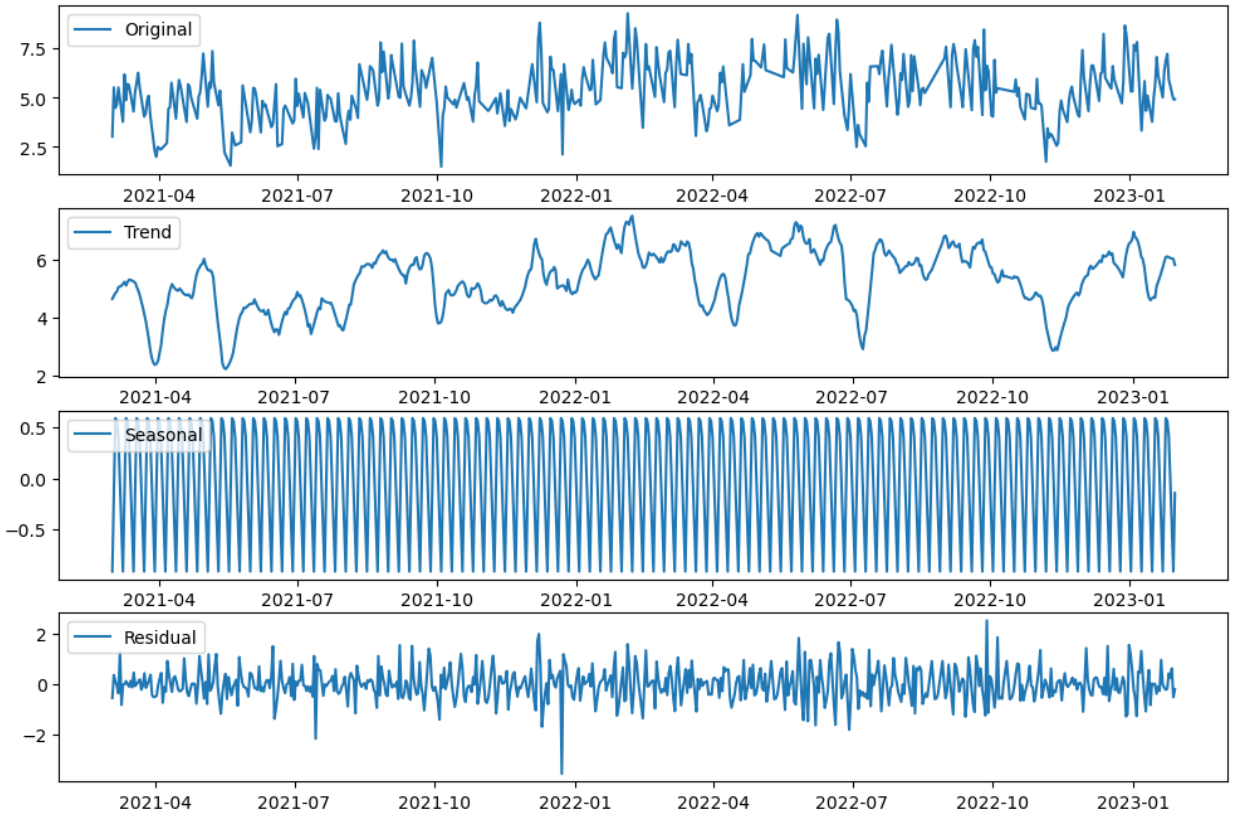


Figure 4.4: seasonal decompose plot for PO_4 in Hias lab dataset

Stationarity test:

The Ad-fuller test(ADF) and KPSS tests, and the ACF and PACF plots were used to verify the stationarity of time series. As discussed in method section if p-value < 0.05 then we can reject the

null hypothesis and consider the series as stationary. The result of ADF test with the p-value=0.05 has been shown below:

```
ADF Statistic: -3.3396386431185765
p-value: 0.059990430803381865
Lags used: 20
Number of observations used: 681
Critical values: {'1%': -3.9721255561961075, '5%': -3.416956617564839, '10%': -3.1308553037306632}
```

Whereas, for KPSS test, p-value should be > 0.05 to reject the null hypothesis. In this case KPSS test returns p-value as 0.01.

```
Kwiatkowski-Phillips-Schmidt-Shin test
------------------------------------------
KPSS Statistic: 1.098246
p-value: 0.010000
Number of lags used: 14
Critical values of KPSS test:
10% 0.347
5% 0.463
2.5% 0.574
1% 0.739
```

From two above tests we can conclude the time series for PO_4 in is not stationary.

ACF and PACF plot:

to ensure that the data for ARIMA analysis is stationary or not, the AC and PAC graphs were plotted and presented below. From this plot we can see that it shows that seasonality is present in the data as it does not converge to zero.

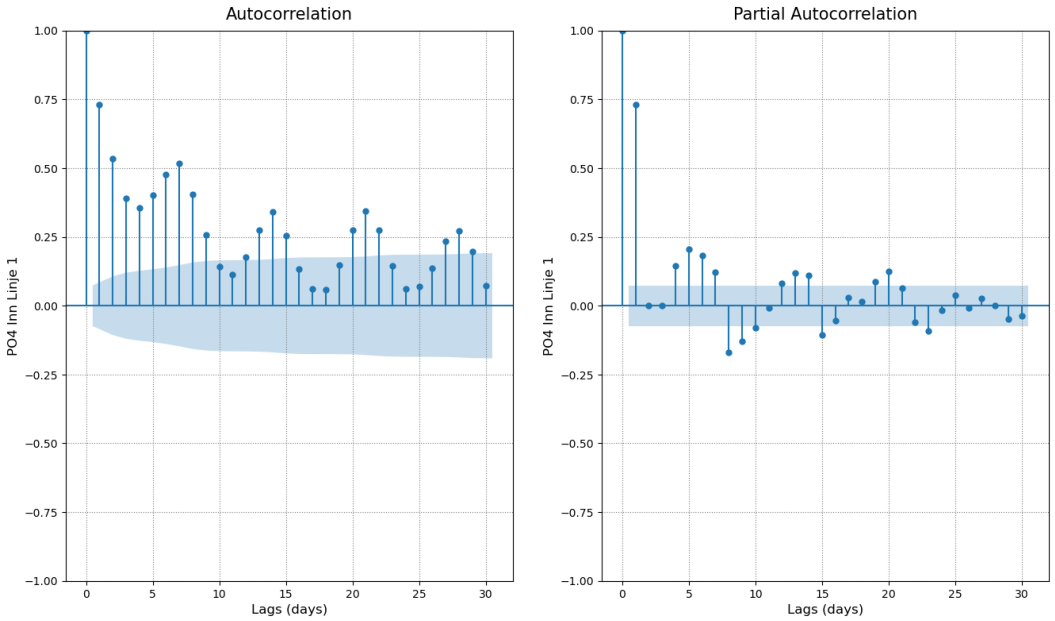


Figure 4.5: ACF / PACF plot for PO_4 in Hias lab dataset

Differentiation: As two previous tests showed that the data is non-stationary, we need to differentiate the time-series to make it stationary before applying to ARIMA model. Below the differentiated time series has been presented in figure 4.6.

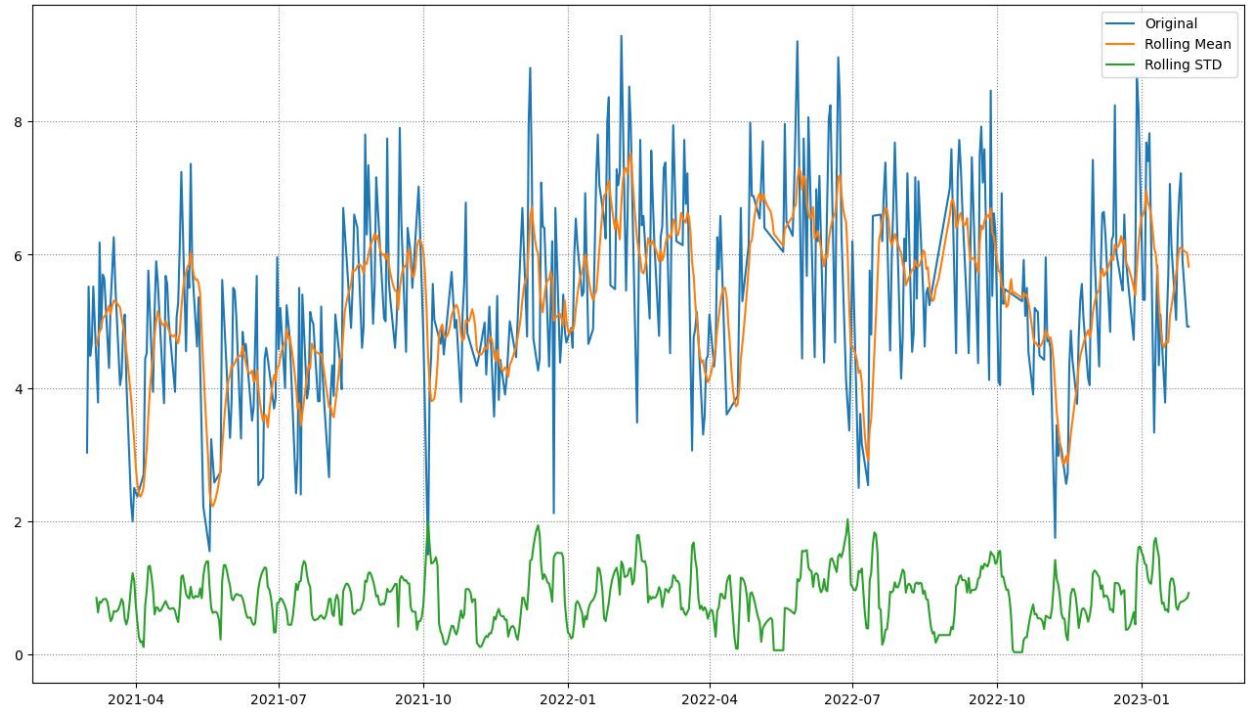


Figure 4.6: Differentiated plot for PO_4 in hias lab dataset

Stage 2. Parameter estimation and model selection

4.1.3 ARIMA model with Hias laboratory dataset:

After differentiation, the stationary time-series is used to calibrate ARIMA time series model for prediction of PO_4 at inlet. The ARIMA(p,d,q) model were ascertained by potential values for the non-seasonal AR order (*p*),nonseasonal MA order (*q*).

Determine the optimum parameters: The selection of the best fitting model from all potential ones was based on the lowest AIC and normalized BIC, Ljung-Box test result and the highest R^2.(Do et al., 2022) This best fitting model finding was done by python pandas library “pmdarima” “auto_arima”.Table 4.1 presents the results of those evaluation metrics for all potential models.

Table 4.1: ARIMA potential models for PO_4 in Hias lab dataset

```
ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=424.341, Time=0.26 sec
ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=515.301, Time=0.01 sec
ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=424.374, Time=0.04 sec
ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=453.146, Time=0.03 sec
ARIMA(0,0,0)(0,0,0)[0]             : AIC=961.806, Time=0.01 sec
ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=426.293, Time=0.18 sec
ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=428.134, Time=0.09 sec
ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=426.412, Time=0.27 sec
ARIMA(2,0,3)(0,0,0)[0] intercept   : AIC=431.020, Time=0.17 sec
ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=426.167, Time=0.04 sec
ARIMA(1,0,3)(0,0,0)[0] intercept   : AIC=421.137, Time=0.22 sec
ARIMA(0,0,3)(0,0,0)[0] intercept   : AIC=428.928, Time=0.05 sec
ARIMA(1,0,4)(0,0,0)[0] intercept   : AIC=423.020, Time=0.27 sec
ARIMA(0,0,2)(0,0,0)[0] intercept   : AIC=431.203, Time=0.04 sec
ARIMA(0,0,4)(0,0,0)[0] intercept   : AIC=427.916, Time=0.07 sec
ARIMA(2,0,4)(0,0,0)[0] intercept   : AIC=413.951, Time=0.26 sec
ARIMA(3,0,4)(0,0,0)[0] intercept   : AIC=383.956, Time=0.33 sec
ARIMA(3,0,3)(0,0,0)[0] intercept   : AIC=382.086, Time=0.27 sec
ARIMA(4,0,3)(0,0,0)[0] intercept   : AIC=426.373, Time=0.32 sec
ARIMA(4,0,2)(0,0,0)[0] intercept   : AIC=390.339, Time=0.30 sec
ARIMA(4,0,4)(0,0,0)[0] intercept   : AIC=385.075, Time=0.33 sec
ARIMA(3,0,3)(0,0,0)[0]             : AIC=388.843, Time=0.24 sec

Best model:  ARIMA(3,0,3)(0,0,0)[0] intercept
Total fit time: 3.792 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	153
Model:	SARIMAX(3, 0, 3)	**Log Likelihood**	-183.043
Date:	Mon, 20 Feb 2023	**AIC**	382.086
Time:	14:30:56	**BIC**	406.330
Sample:	0	**HQIC**	391.934
	- 153		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.2929	0.200	1.464	0.143	-0.099	0.685
ar.L1	2.1702	0.054	39.989	0.000	2.064	2.277
ar.L2	-2.1402	0.070	-30.377	0.000	-2.278	-2.002
ar.L3	0.9167	0.051	17.993	0.000	0.817	1.017
ma.L1	-1.7071	0.107	-15.961	0.000	-1.917	-1.497
ma.L2	1.4689	0.157	9.341	0.000	1.161	1.777
ma.L3	-0.4228	0.116	-3.640	0.000	-0.650	-0.195
sigma2	0.6257	0.059	10.578	0.000	0.510	0.742

Ljung-Box (L1) (Q):	0.02	**Jarque-Bera (JB):**	39.25
Prob(Q):	0.88	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.67	**Skew:**	0.38
Prob(H) (two-sided):	0.15	**Kurtosis:**	5.36

Stage 3. Forecasting

Wastewater inlet PO_4 concentration forecasting: The ability of the proposed ARIMA model in predicting wastewater inlet PO_4 concentration data was assessed in this last stage. The testing dataset (2 January to 31 january 2023) was used for the model validation procedure. The ARIMA(3,0,3) model was directly utilized for the entire testing process. There were 2 weeks (1st February to 14th February 2023) forecasts generated outside the testing period, as this study mainly focuses on the predicting future values. The result of model calibration showing a comparison between the model predicted and lab-analysed values is presented in Figure 4.7. Two weeks forecast of inlet PO_4 concentration plot is shown in figure 4.8. ARIMA model was evaluated with statistical analysis and result was not satisfactory with RMSE 1.45 and coefficient of determination (R^2) = 0.72.

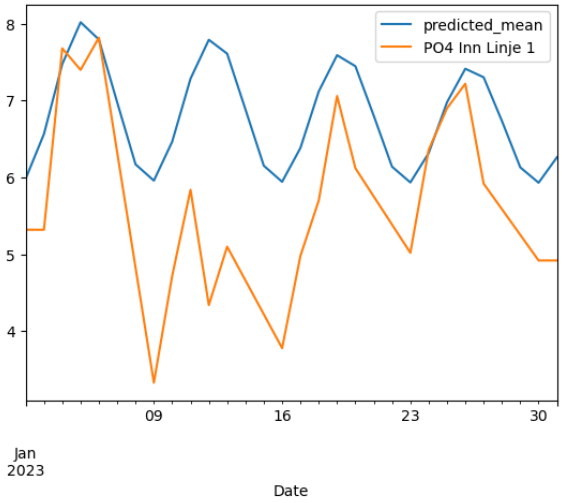


Figure 4.7: Model evaluation with comparison of predicted mean and testing dataset

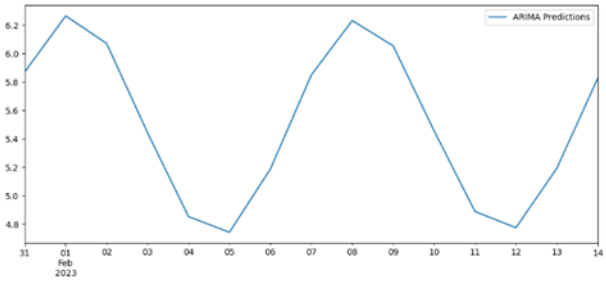


Figure 4.8: two weeks advance prediction of PO_4 in

4.1.4 SARIMAX Model with Hias laboratory dataset:

The same differentiated time-series of PO_4 in has been used to develop SARIMAX model for PO_4 prediction with input sCOD, Flow and NH_4 as Exogeneous inputs.

Table 4.2: Best fit SARIMAX model for prediction of PO_4

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	702
Model:	SARIMAX(1, 1, 1)	**Log Likelihood**	-733.374
Date:	Thu, 27 Apr 2023	**AIC**	1476.747
Time:	16:29:38	**BIC**	1499.510
Sample:	03-01-2021	**HQIC**	1485.546
	- 01-31-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
NH4 inn	0.0287	0.003	11.276	0.000	0.024	0.034
SCOD Inn	0.0049	0.000	19.206	0.000	0.004	0.005
ar.L1	0.5168	0.028	18.360	0.000	0.462	0.572
ma.L1	-0.9331	0.016	-57.429	0.000	-0.965	-0.901
sigma2	0.4738	0.015	30.630	0.000	0.443	0.504

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	610.98
Prob(Q):	0.99	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.28	**Skew:**	0.29
Prob(H) (two-sided):	0.06	**Kurtosis:**	7.54

```
                           SARIMAX Results
==============================================================================
Dep. Variable:        PO4 Inn Linje 1   No. Observations:                  561
Model:               SARIMAX(1, 1, 1)   Log Likelihood                -586.741
Date:                Thu, 27 Apr 2023   AIC                           1183.482
Time:                        16:30:24   BIC                           1205.122
Sample:                  03-01-2021     HQIC                          1191.932
                       - 09-12-2022
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
NH4 inn        0.0308      0.003     10.271      0.000       0.025       0.037
SCOD    Inn    0.0047      0.000     15.440      0.000       0.004       0.005
ar.L1          0.4504      0.033     13.597      0.000       0.385       0.515
ma.L1         -0.9142      0.021    -42.923      0.000      -0.956      -0.872
sigma2         0.4752      0.018     26.905      0.000       0.441       0.510
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               517.20
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               1.63   Skew:                             0.12
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.70
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The Best fit model has been chosen with auto-arima function and it shows the result that best model was SARIMAX(1, 1, 1) (0,0,0)7 where seasonal P, D,Q are 0 as the series has no seasonlaity and s=7 as the series has daily data frequency.

Below the result of diagonstic test has been presented in figure 4.9 . Residual plot has not been converged near zero and the predicted plot was unable to catch the trend of actual test dataset, which is shown in figure . The Model evaluation score was not satisfacory with $R^2 = 0.43$.

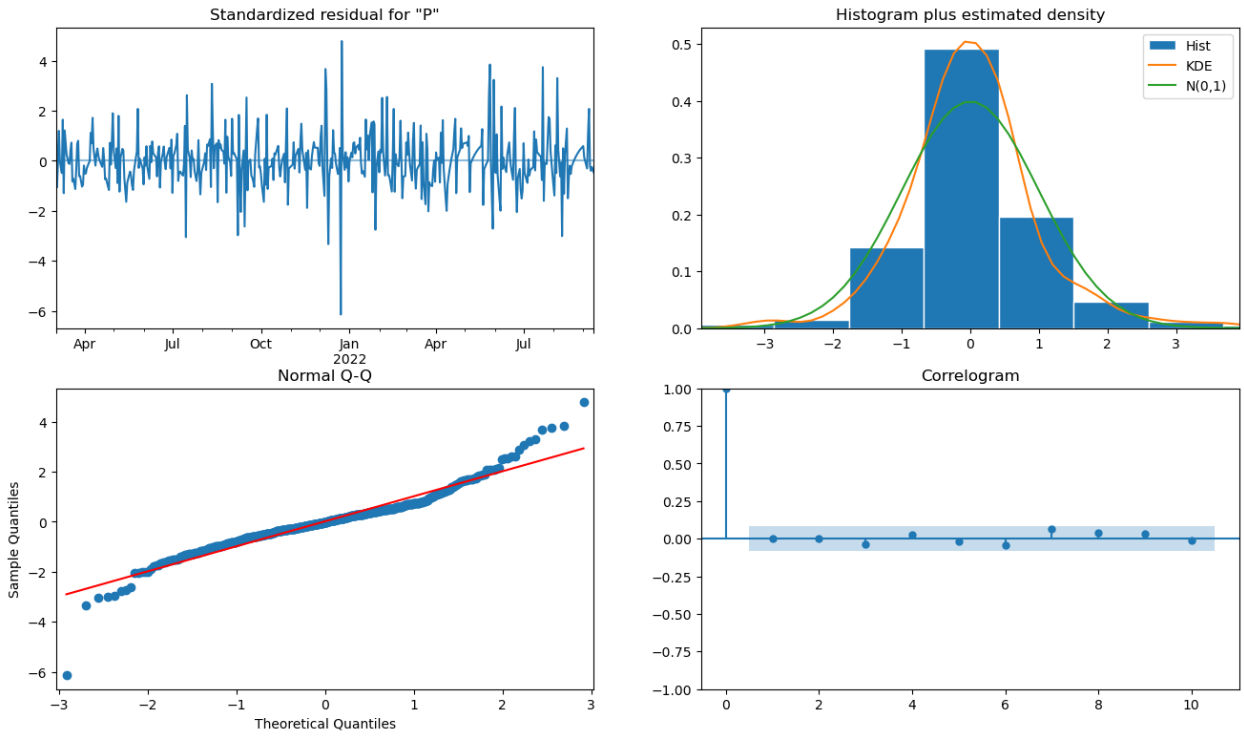


Figure 4.9: Diagnostic check for SARIMAX model For PO_4 prediction

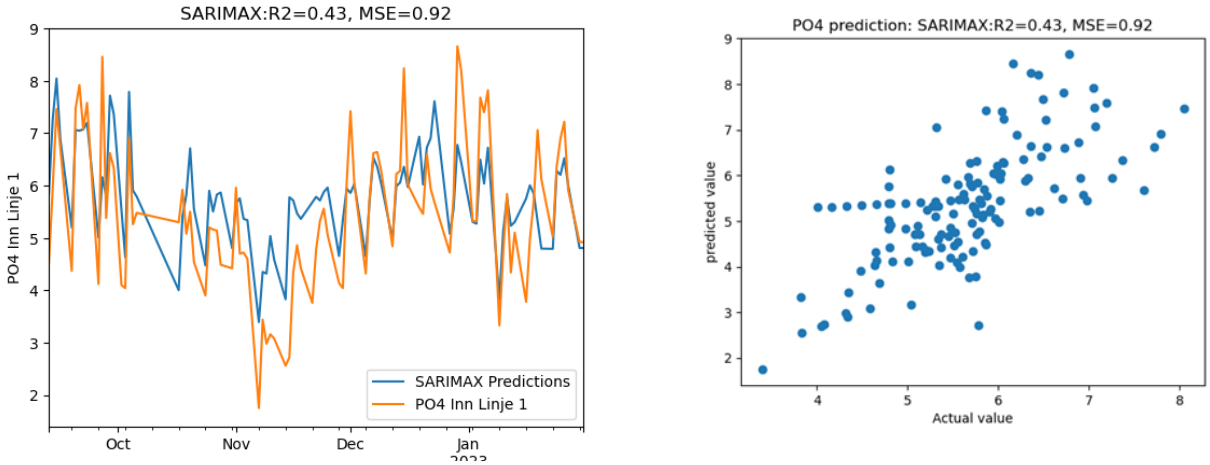


Figure 4.10: Predicted vs actual test test dataset plot for PO_4 prediction with SARIMAX model

In addition, we have tried SARIMAX model for sCOD prediction with input Flow PO_4in and NH_4 and the result of best model fit, and diagnostic check have been presented in appendix section A in figure A.1 and figure A.2 respectively. The best model was found as SARIMAX (1, 1,,4)(1,0, 1)7 where seasonal D=0 and s=7 for daily sampled data. The Result for sCOD prediction with SARIMAX was nearly not satisfactory same as PO_4 prediction with R^2 score as 0.43. The predicted plot vs actual test dataset plot has been presented in below figure.

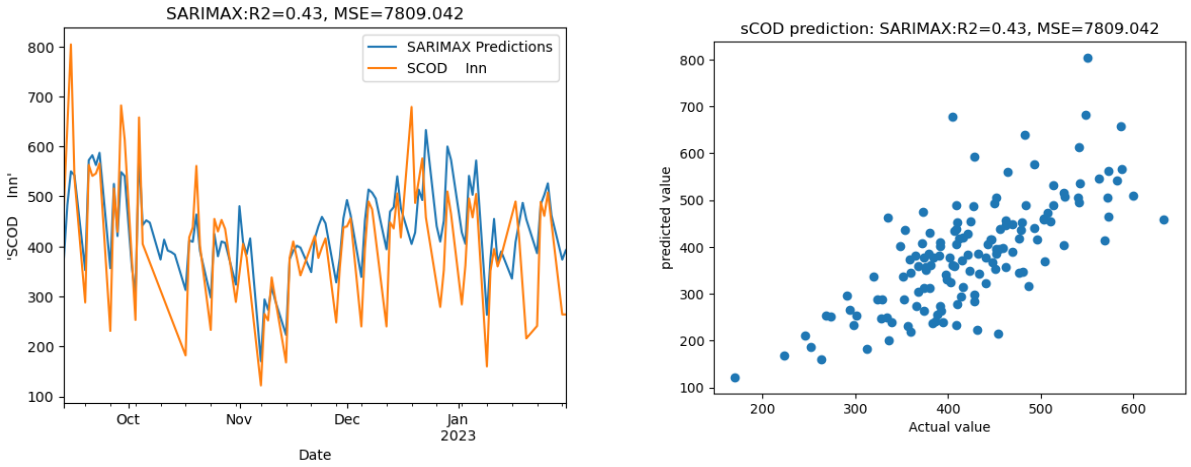


Figure 4.11: Predicted vs. actual plot for sCOD prediction SARIMAX model with Hias lab dataset

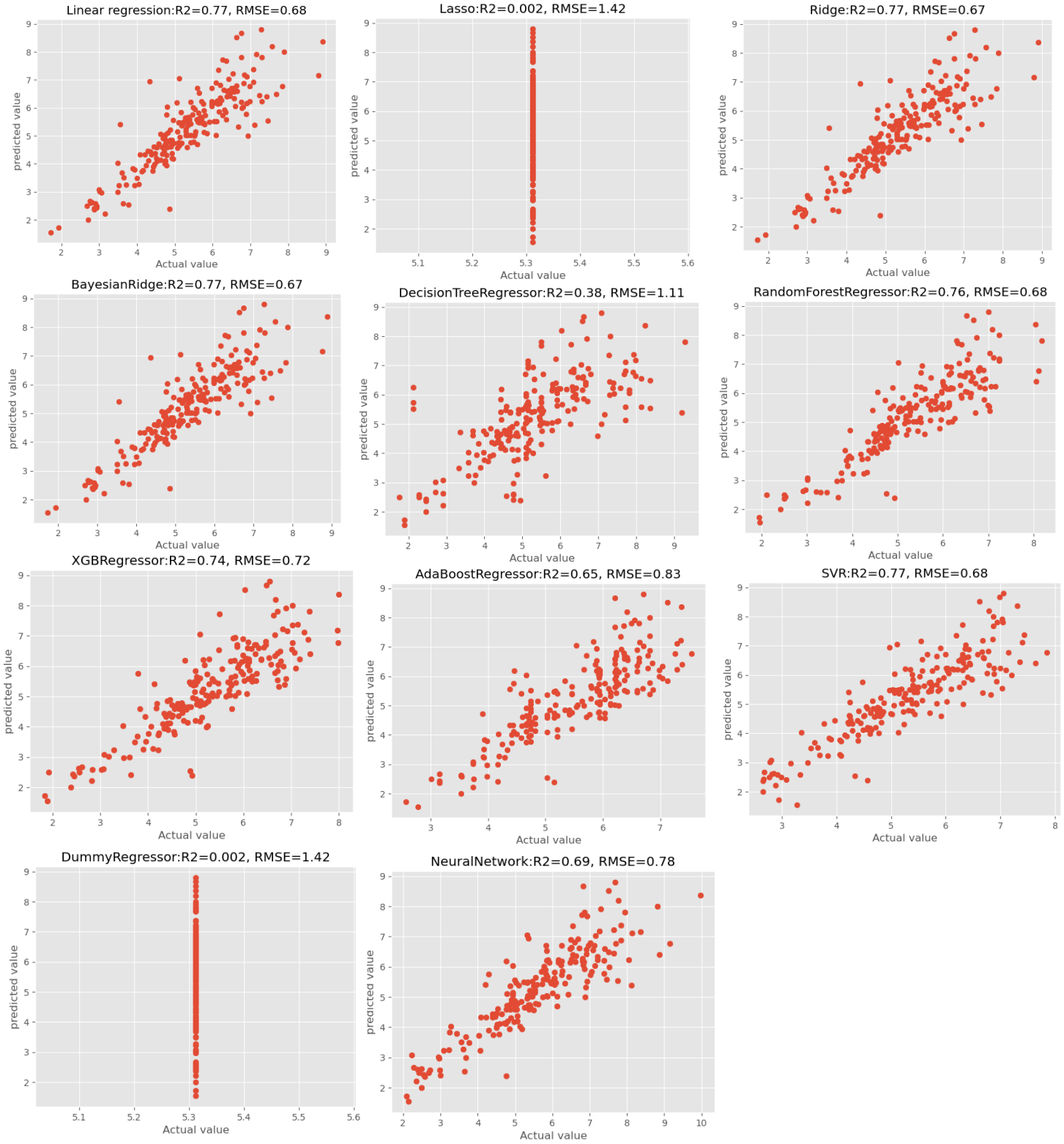
4.1.5. Regression and Neural network with Hias laboratory dataset:

In this study, different machine learning methods were used to predict the influent PO_4 concentration with Flow rate, sCOD in and NH_4-N as input in 2023 in Hias WWRF with different regression analysis and Neural network model.

As regression model has few parameters and model was too simple to capture the non-linearity of data, previous time steps of output and all input have been added to increase the accuracy of all machine learning model. In this way, the accuracy measurement of R^2 value has been increased from 0.62 to 0.77 in case of linear regression and all other models which is quite significant. K-fold cross validation has been used with K fold value as 5 and number of epochs used in NN is 100. The function first takes in training and test data along with a list of algorithms to test. It trains each algorithm using cross-validation, and then uses the best algorithm on the test data to

evaluate its performance. It returns a summary data frame containing evaluation metrics (R^2, mae, rmse) for each algorithm and a list of predicted y values for the best algorithm.

The results of model calibration showing a comparison between the model predicted and lab measured values are presented in Figure 11. The plots also show the root mean-square error (RMSE) and the coefficient of determination (R^2) for the models calibrated using the different calibration algorithms. The scatter plot of a perfect prediction model would be a 45-degree line with a slope = 1 and intercept = 0 and an R^2 value of 1. A quantitative assessment of different prediction models is conducted by comparing the values of R^2 and regression line equations. The plots presented in Figure 4.12 as well as a comparison between correlation coefficients (R^2 and MSE) show minimal difference between the results obtained with all the algorithms. However, the model coefficient obtained by the SVR, Linear Regression algorithm, Ridge and Bayesian Ridge algorithm shows relatively better results with (R^2 =0.77 and RMSE = 0.68) as compared to an R^2 =0.68 for Ada Boost regression regression algorithm, R^2 =0.38 for Decision Tree Regression algorithm. Lasso algorithm cannot capture the learning trend and showed extremely poor result. Neural Network showed satisfactory results with R^2 as 0.69 and RMSE as 0.78.



Figur 4.22: Lab-measured versus predicted values using different model calibration algorithms for Hias Lab analyzed data.

4.1.6. Model evaluation:

The result of model performance evaluation for all models for influent parameter prediction with HIAS Lab-dataset have been tabulated in below table:

Table 4.3: Evaluation of different model with HIAS Lab - dataset

Model	**Predicted parameters**	**Input parameters**	**R^2**	**RMSE**
ARIMA	PO_4	Previous values of PO_4	0.72	1.45
SARIMAX	PO_4	Flow rate, sCOD, NH_4-N	0.43	0.96
	sCOD	NH_4-N, PO_4, Flow rate	0.42	88.3
LinearRegression	PO_4	Flow rate, sCOD, NH_4-N	0.77	0.68
Lasso			0.002	1.42
Ridge			0.77	0.67
BayesianRidge			0.77	0.67
DecisionTreeRegressor			0.38	1.11
RandomForestRegressor			0.76	0.68
XGBRegressor			0.74	0.72
AdaBoostRegressor			0.65	0.83
SVR			0.77	0.68
DummyRegressor			0.002	1.42
NeuralNetwork			0.69	0.78

4.2. Result HIAS online-data

4.2.1. Result of data analysis

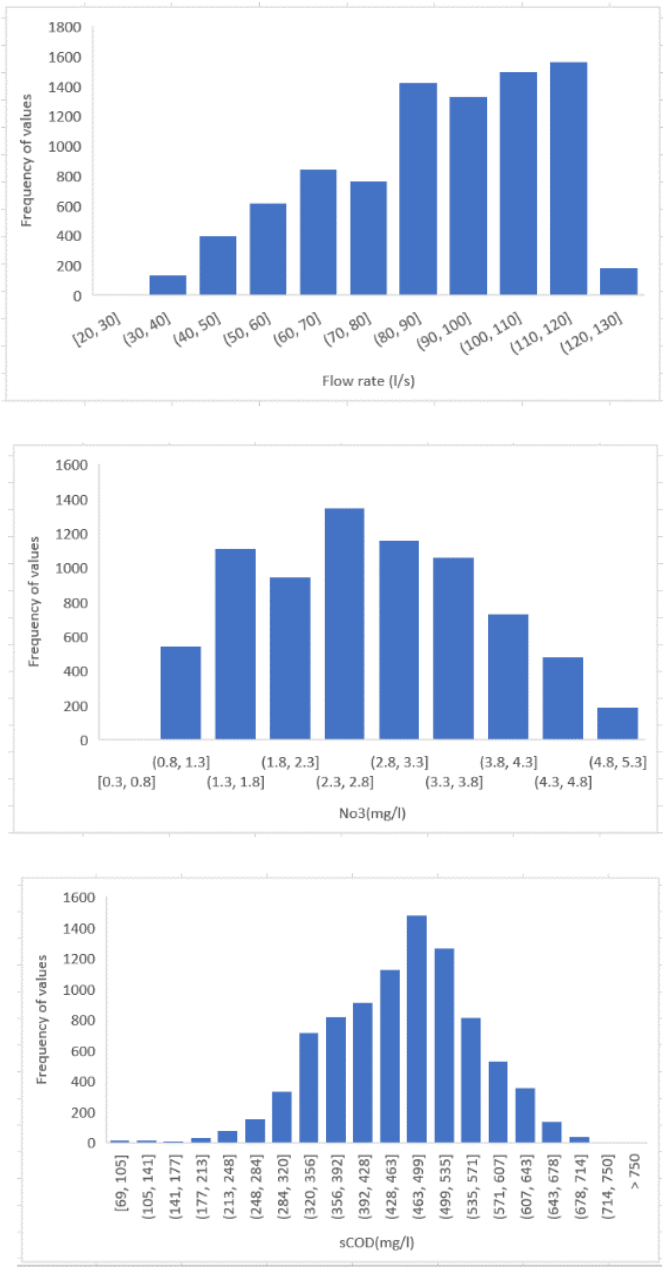


Figure (4.13a): Distribution of data from December2022-January2023 from online measurement (Hias)

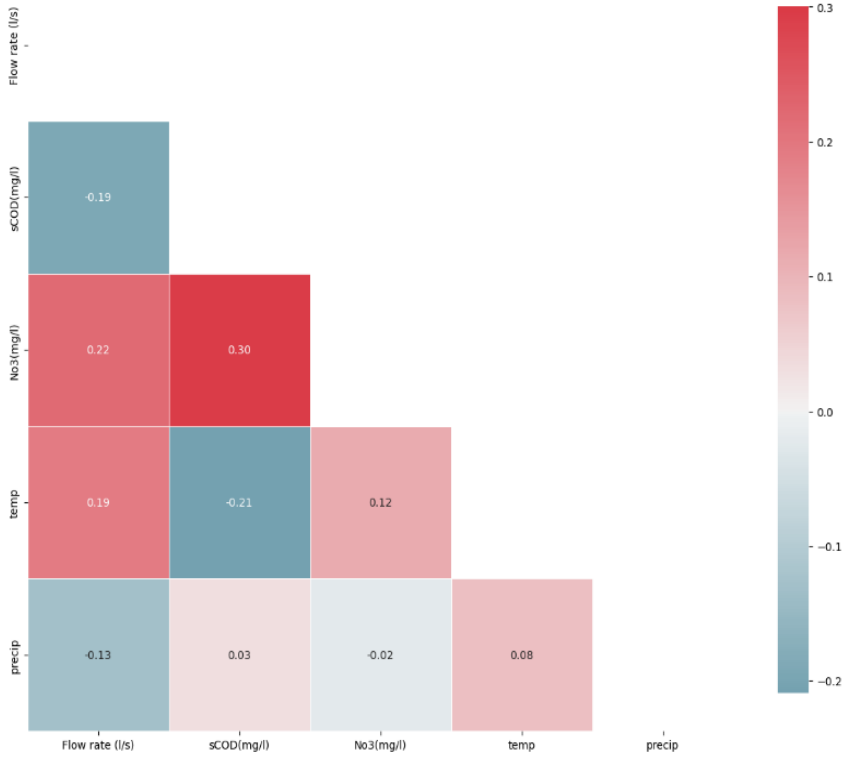


Figure 4.13b: Heatmap of correlation between parameters of online dataset (Hias)

The correlations among wastewater-influent parameters collected for Hias online dataset is illustrated in Fig. 4.13b. Same as lab-dataset flow rate and sCOD is negatively correlated with coefficient -0.19, whereas NO_3 is positively correlated with sCOD with correlation coefficient 0.30. Apart from these, correlation matrix demonstrates very poor correlation between different parameters. The results indicated that advanced algorithms are needed for predicting inlet sCOD, rather than basic statistical methods. In general, there are two typical approaches to predicting the output variable; In the first approach, the output variable is constructed based on other input variables (SARIMAX, MLR), while in the second, the output variable relies on the previous output variable data itself (ARIMA, SARIMA). (Ly et al., 2022)This weak correlation coefficient suggests that the second approach is a better option for predicting inlet sCOD or flow rate with online dataset.

Analysis of Average temperature and Precipitation

Historical data of average temperature and precipitation in Hamar, Norway from year 2022 to 2023 is captured from yr.no. The aim to collect this dataset was to use it as exogeneous dataset. As correlation matrix (figure 4.13b) showed negative correlation between precipitation and flow rate which is not meaningful, we need to proceed without average temperature and precipitation parameters for model development.

After data preprocessing and removing all outliers cleaned data of sCOD(mg/l), flow rate(l/s) and NO_3(mg/l) have been chosen as parameters for further model development.

4.2.2. Result of Model Development

Stage 1. Model identification

Same as Hias lab analyzed data, the stationarity of Hias online dataset has been checked original training time series of inlet sCOD concentration with 10min sampling frequency was investigated by plotting and decomposing the time series into its components to observe any trends, seasonality, and residuals. Here the period of sesaonal_decompose function was chosen 144 as data was sampled 10 min frequency(24*6=144).

decompose_data = seasonal_decompose(data_ten_min['sCOD(mg/l)'], period=144)

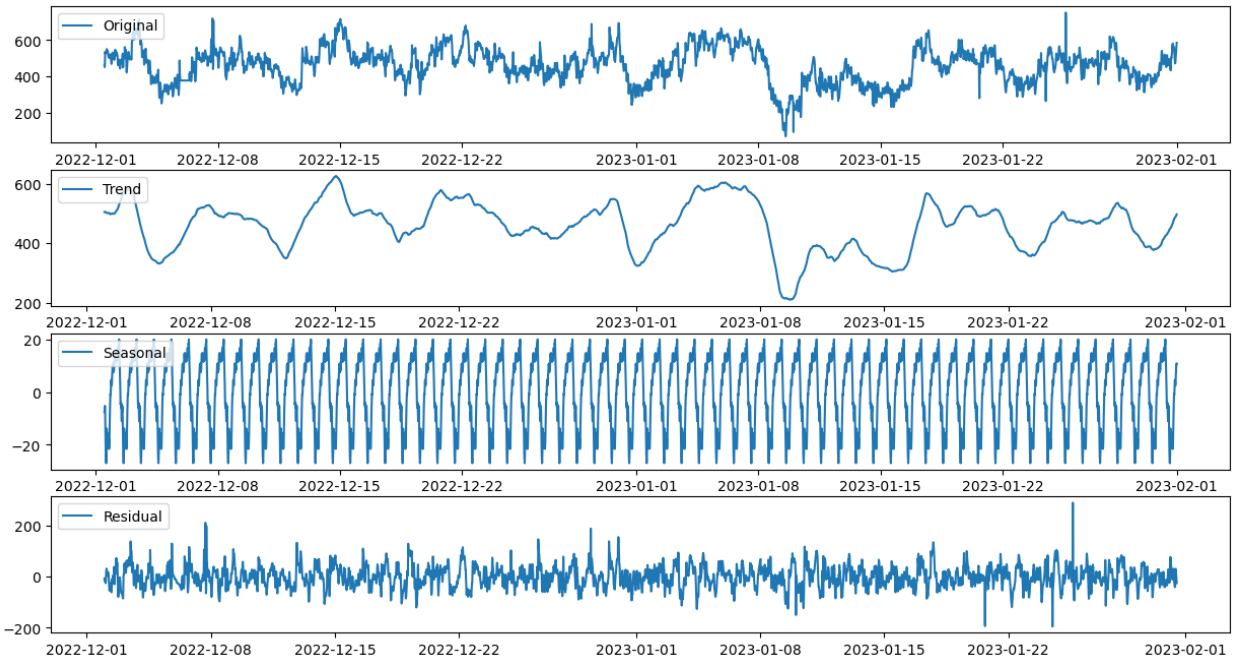


Figure 4.14: seasonal decompose plot for sCOD of online dataset (Hias)

Stationarity test:

The Ad-fuller test(ADF) and KPSS tests, and the ACF and PACF plots were further used to verify the data's stationarity.

```
ADF Statistic: -5.608364218299525
p-value: 1.4020833656437163e-05
Lags used: 25
Number of observations used: 8793
Critical values: {'1%': -3.9597999482271584, '5%': -3.410989423202829, '10
%': -3.127344102884938}
```

ADF test return P-value quite < 0.05, so we can reject null hypothesis and conclude the series as stationary.

```
Kwiatkowski-Phillips-Schmidt-Shin test
------------------------------------------
KPSS Statistic: 0.735242
p-value: 0.010342
Number of lags used: 56
Critical values of KPSS test:
10% 0.347
5% 0.463
2.5% 0.574
1% 0.739
```

Whereas, from KPSS test result we can see the series is non-stationary. For this contradiction of result, the time series was further tested with ACF/PACF test.

ACF and PACF plot:

To ensure that the data for ARIMA analysis is stationary, the AC and PAC graphs were analyzed with Hias online data and presented below. From this plot we can see both the AC and PAC graphs

do not meet all the criteria. They tended to show a seasonality, and most of the correlations does not remain within the confidence limits.(Boyd et al., 2019)

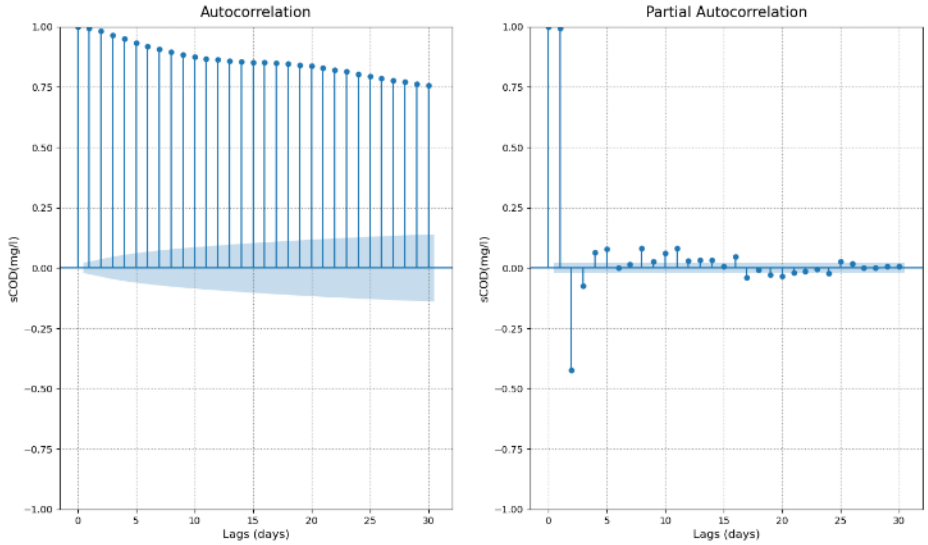


Figure 4.15: ACF/PACF plot for origianl time series sCOD of online dataset(Hias).

Differentiation

As above two test show the presence of seasonality in the series, we have twice differentiated the series to make it stationary. The ACF plot for two times differentiation has been shown in below figures.

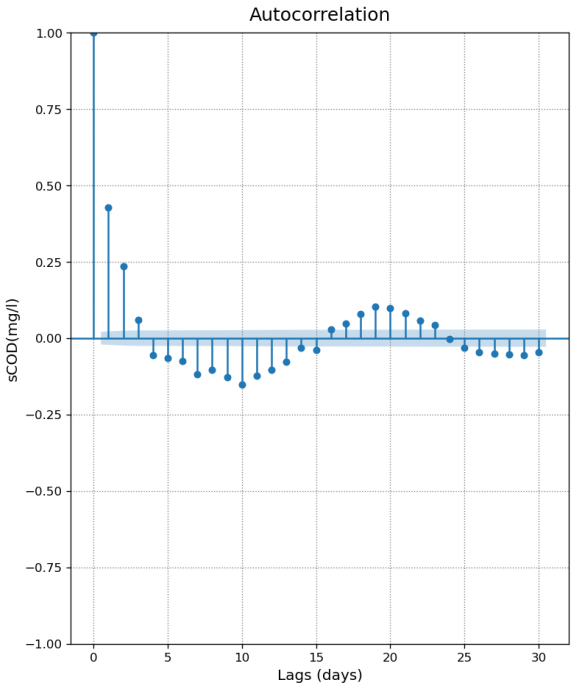


Figure 4.16a: ACF plot of 1st order differentiation of sCOD

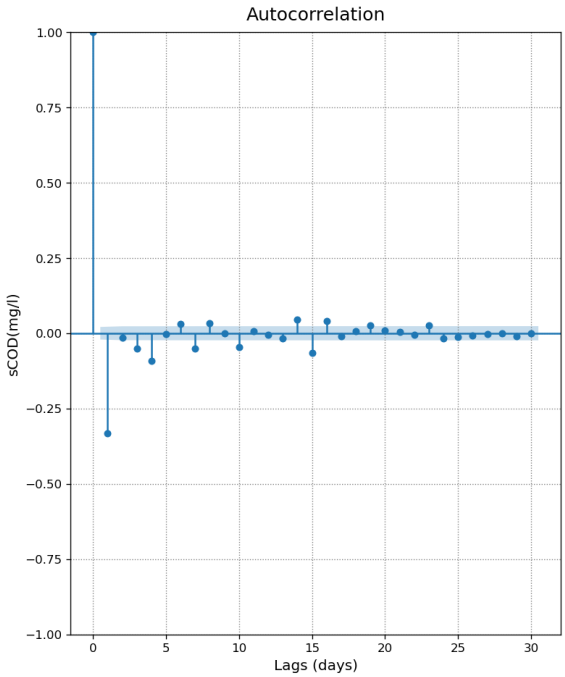


Figure 4.16b: ACF plot of 2nd order differentiation of sCOD

Stage 2. Parameter estimation and model selection

4.2.3 ARIMA model with Hias online dataset:

As it is very poor correlation coefficient between different parameters of Hias online dataset, this dataset is used to calibrate ARIMA time series model for prediction of sCOD at inlet.

The ARIMA(p,d,q) model were ascertained by potential values for the non-seasonal AR order (*p*),nonseasonal MA order (*q*).

Determine the optimum parameters: The selection of the best fitting model was done by python pandas library “pmdarima” “auto_arima”. Figure A.3 in appendix section(A) presents the results of those evaluation metrics for all potential models. The best fit model was ARIMA (0,0,1) where both non seasonal p and d terms were 0.

Stage 3. Forecasting

Wastewater inlet sCOD concentration forecasting: The ability of the proposed ARIMA model in predicting wastewater inlet sCOD concentration data was assessed in this stage. The testing dataset (20% of totat dataset) was used for the model validation procedure. The ARIMA (0,0,1) model was directly utilized for the entire testing process. The result of model calibration showing a comparison between the model predicted and test dataset values is presented in Figure 4.17. RMSE score was 11.5 and R^2 score was 0.21.

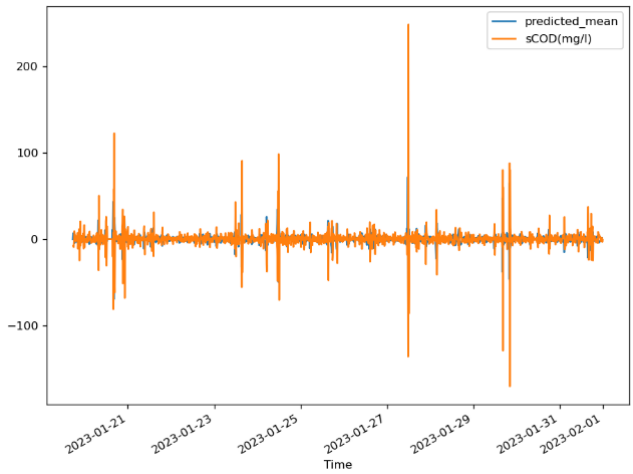


Figure 4.17: Model evaluation with comparison of predicted mean and testing dataset for sCOD prediction with online dataset(Hias)

4.1.5. Regression and Neural network with Hias online dataset:

In this study, different machine learning methods were used to predict the influent SCOD concentration from HIAS online dataset.

The results of model calibration showing a comparison between the model predicted and actual values are presented in Figure 4.18. The table below also show the root mean-square error (RMSE) and The coefficient of determination (R-squared (R^2) for the models calibrated using the different calibration algorithms. The plots presented in Figure 4.18 as well as a comparison between correlation coefficients (R^2 and MSE) show minimal difference between the results obtained with all the algorithms. However, the model coefficient obtained by the Linear regression, Ridge, Lasso, Random forest and Neural network algorithm shows relatively better

results with (R^2 =0.891) as compared to an R^2 =0.63 for SVR algorithm, R^2 =0.76 for decision tree algorithm.

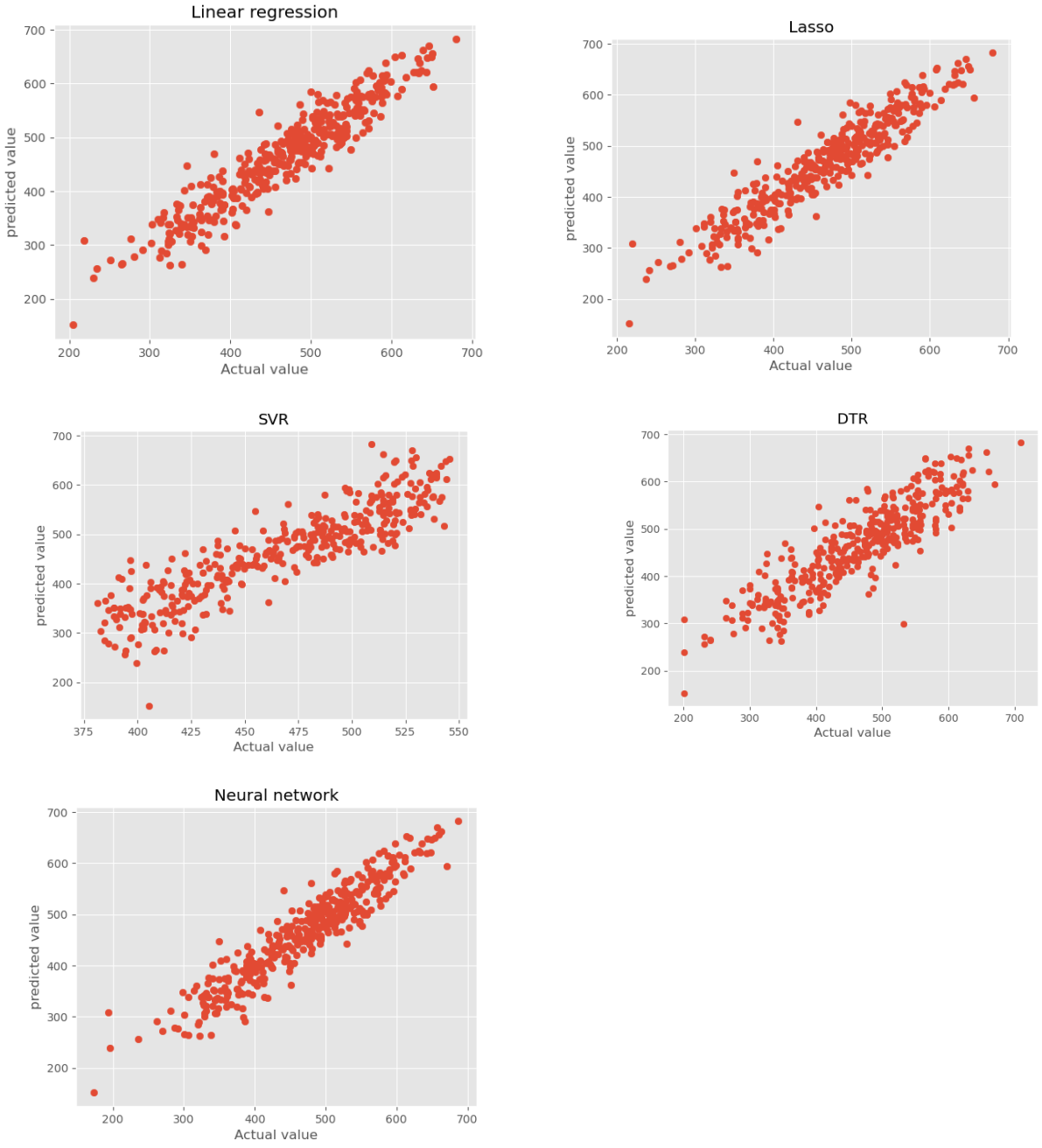


Figure 4.18: Predicted vs. Actual plot for different regression model and Neural network for sCOD in online dataset(Hias).

4.2.6. Model evaluation:

Below table has shown summary of model performance evaluation for all models for prediction of SCOD in with HIAS online-dataset

Table 4.4: Evaluation of different model with HIAS online dataset

Model	Predicted parameters	Input parameters	R^2	RMSE
ARIMA	sCOD	Previous values of sCOD	0.21	11.5
LinearRegression	sCOD	Flow rate, NO_3	0.89	30.6
Lasso			0.88	31.20
Ridge			0.89	30.6
BayesianRidge			0.89	30.6
DecisionTreeRegressor			0.76	45.6
RandomForestRegressor			0.88	32.01
XGBRegressor			0.86	34.66
AdaBoostRegressor			0.86	34.50
SVR			0.63	56.17
DummyRegressor			0.002	93.24
NeuralNetwork			0.88	31.66

4.3. Result VEAS Lab-analyzed data

4.3.1. Result of data analysis

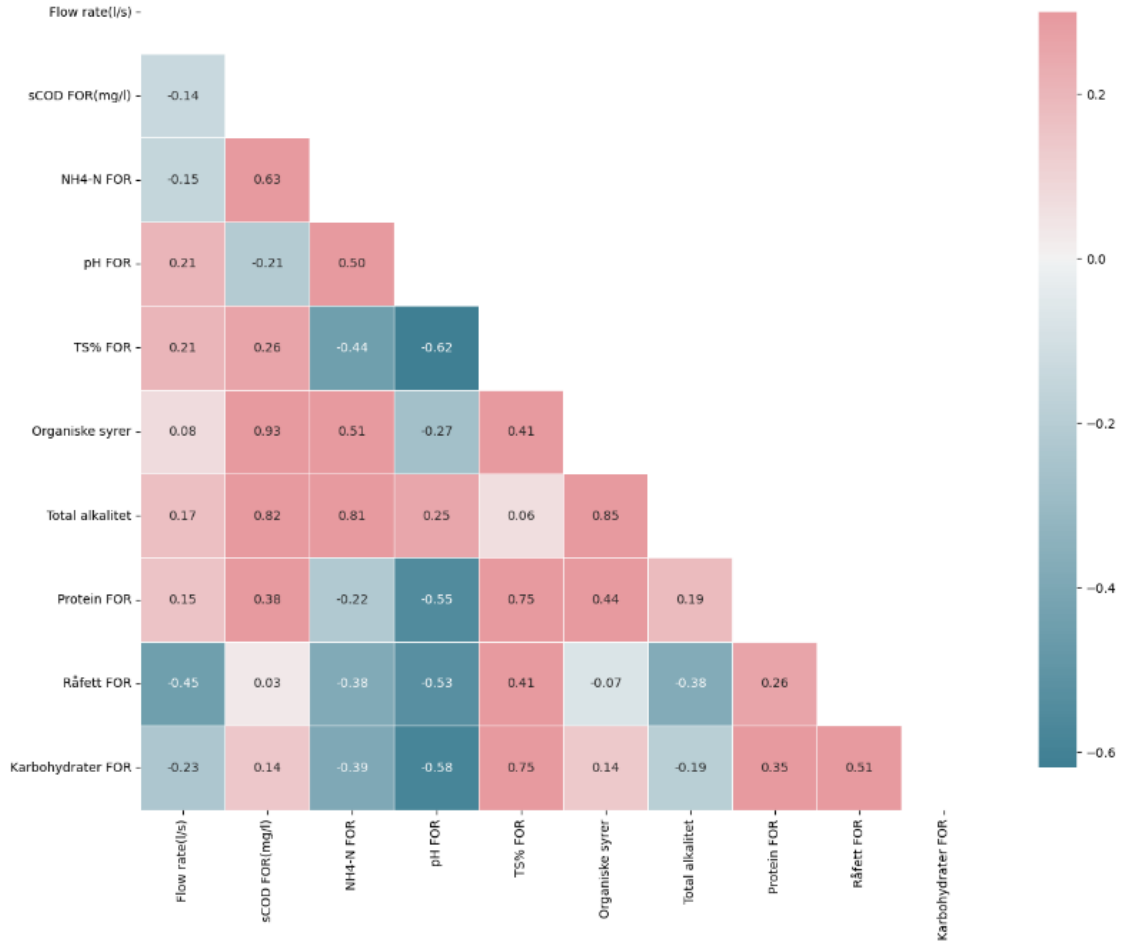


Figure 4.19: Correlation heatmap matrix between different parameters from VEAS lab dataset

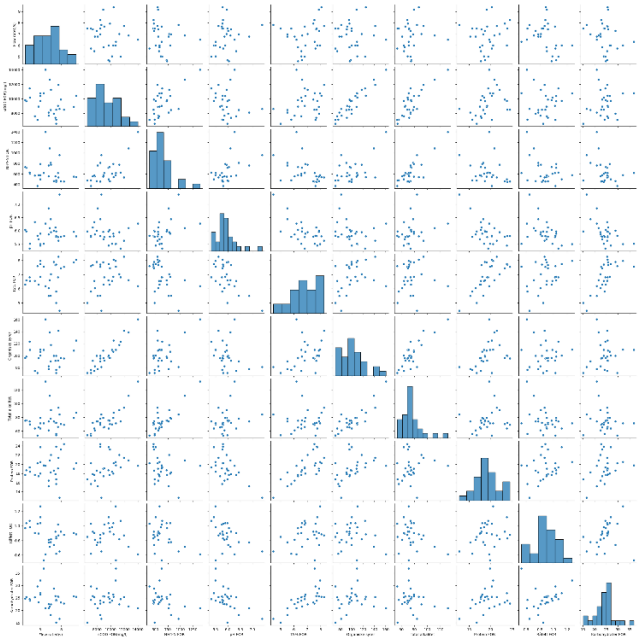


Figure 4.20: Pairplot between different parameters from VEAS lab dataset

The correlations among sludge-influent parameters collected for VEAS laboratory-analyzed data is illustrated in Fig. 4.19. Organic acid or fatty acid is an important compound of sludge influent composition, which is the result of the decomposition of organic matter during the wastewater treatment process. Organic acids can have a negative impact on the quality of the treated wastewater if they are not properly managed. High concentrations of this compound can lead to increased biological oxygen demand (BOD) and chemical oxygen demand (COD) in the treated wastewater, which can be visualized by positive correlation with high coefficient (0.93) in figure 8. Due to corrosive and complex nature of sludge components, it is difficult to measure organic acid with sensor, which leads us to develop **organic acid prediction model with flow rate, NH_4-N, pH and TS% as input hourly data.**

4.3.2. Result of Model Development

Stage 1. Model identification

Same as before, the stationarity of VEAS lab analyzede dataset has been checked original training time series of inlet Organic acid concentration with 1hour sampling frequency was investigated by plotting and decomposing the time series into its components to observe any trends,

seasonality, and residuals. Here the period of sesaonal_decompose function was chosen 24 as data was sampled hourly.

decompose_data = seasonal_decompose(df_hourly['Organiske syrer'], period=24)

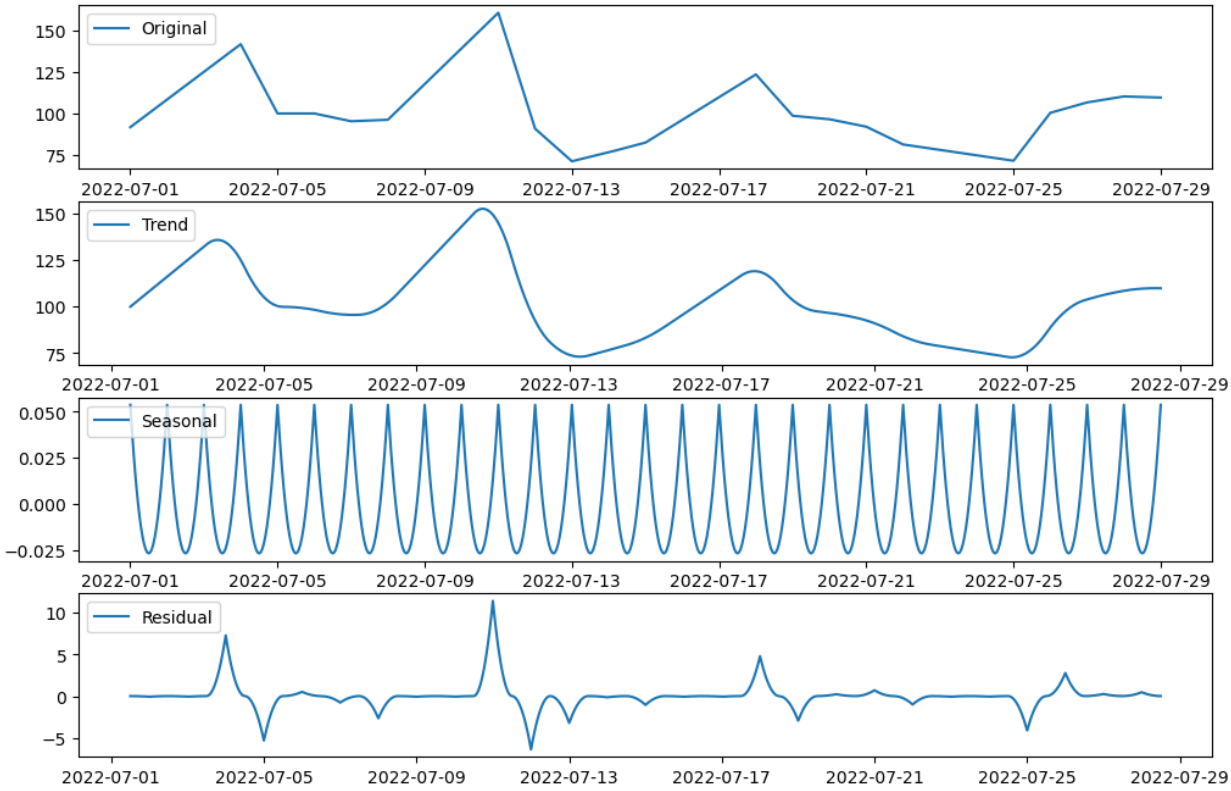


Figure 4.21: seasonal decompose plot for organic acid in VEAS lab dataset

Stationarity test:

The Ad-fuller test(ADF) and KPSS tests, and the ACF and PACF plots were further used to verify the data's stationarity.

```
ADF Statistic: -4.239825386616874
p-value: 0.0038948926384970305
Lags used: 1
Number of observations used: 671
Critical values: {'1%': -3.972325535886194, '5%': -3.4170532894859873, '10%': -3.1309121448362194}
```

ADF test return P-value 0.003 < 0.05, so we can reject null hypothesis and conclude the series as stationary.

```
Kwiatkowski-Phillips-Schmidt-Shin test
```

```
-------------------------------------------
KPSS Statistic: 0.741382
p-value: 0.010000
Number of lags used: 17
Critical values of KPSS test:
10% 0.347
5% 0.463
2.5% 0.574
1% 0.739
```

Whereas, from KPSS test result we can see the series is non-stationary. For this contradiction of result, the time series was further tested with ACF/PACF test.

ACF and PACF plot:

To ensure that the data for ARIMA analysis is stationary, the AC and PAC graphs were analyzed with VEAS lab data and presented below.

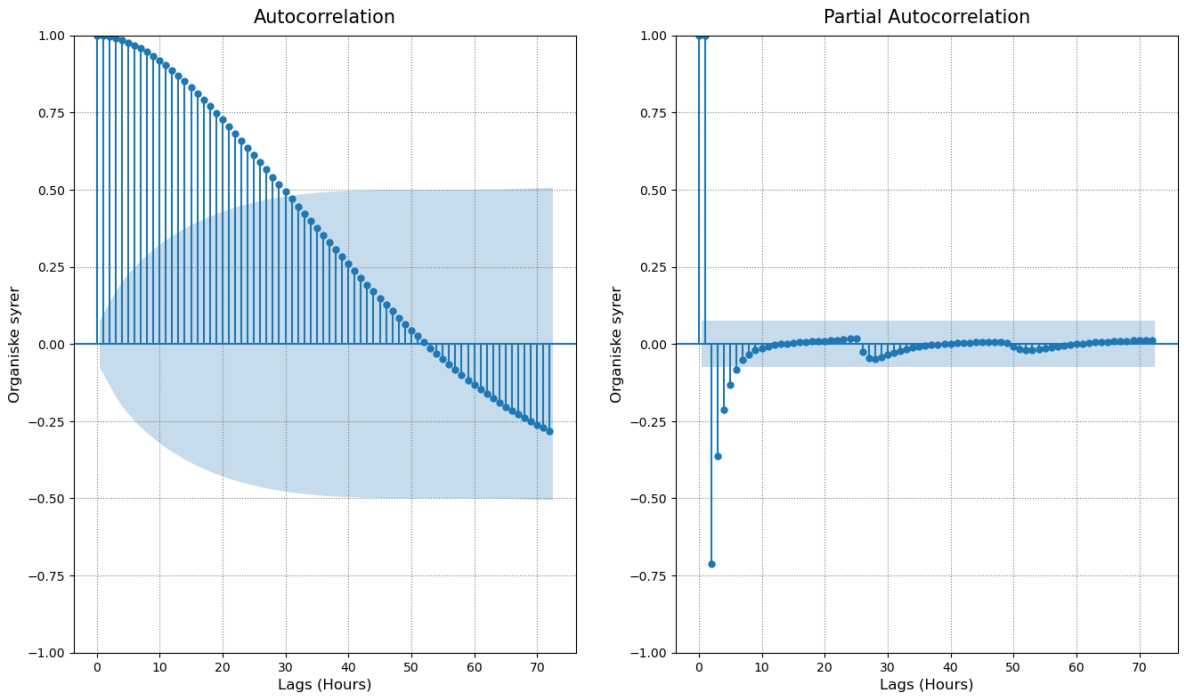


Figure4.22 : Figure 15: ACF/PACF plot for original time series of organic acid in VEAS lab dataset

Stage 2. Parameter estimation and model selection

4.3.3 SARIMAX model with VEAS lab analyzed dataset:

As it was quite good poor correlation coefficient between different parameters of VEAS lab-dataset, this dataset is used to calibrate SARIMAX time series model for prediction of organic acid at inlet with 'Flow rate(l/s)','NH_4-N FOR', 'pH FOR', 'TS% FOR' as exogeneous inputs.

The SARIMAX(p,d,q) (P,D,Q)s model were ascertained by potential values for the non-seasonal and seasonal AR order (*p*) (P) and , MA order (*q*) and (Q)and differentiation order (d), (D) and period for seasonality s .

Determine the optimum parameters: The selection of the best fitting model was done by python pandas' library "pmdarima" "auto_arima". Figure A4 in appendix section presents the results of those evaluation metrics for all potential models. The best fit model was SARIMAX (1,1,1) (2,0,0)24 where both seasonal D and Q terms were 0 and seasonality was s=24.Diagnostic check have been done which shows some deviation from 0 in residual plot.

Stage 3. Forecasting

Sludge inlet Organic acid concentration forecasting: The ability of the proposed SARIMAX model in predicting sludge influent organic acid concentration data was assessed in this stage. The testing dataset (20% of total dataset) was used for the model validation procedure. The result of model calibration showing a comparison between the model predicted and test dataset values is presented in Figure 9. The result was satisfactory with RMSE score being 3.40 and R^2 score was 0.95.

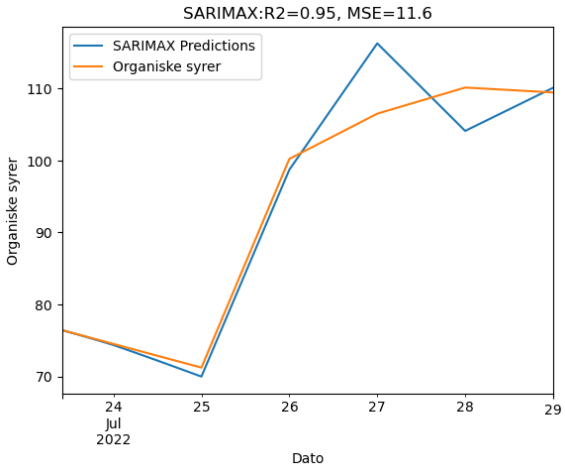


Figure 4.23: Model evaluation with comparison of predicted mean and testing dataset SARIMAX for organic acid in Veas lab dataset

4.3.4. Regression and Neural network with VEAS lab analyzed dataset:

In this study, different machine learning methods were used to predict Sludge inlet Organic acid concentration from VEAS lab dataset.

The results of model calibration showing a comparison between the model predicted and actual values are presented in Figure 4.24. The table below also show the root mean-square error (RMSE) and the coefficient of determination (R-squared (R^2) for the models calibrated using the different calibration algorithms. The plots presented in Figure 4.24 as well as a comparison between correlation coefficients (R^2 and MSE) show minimal difference between the results obtained with all the algorithms.

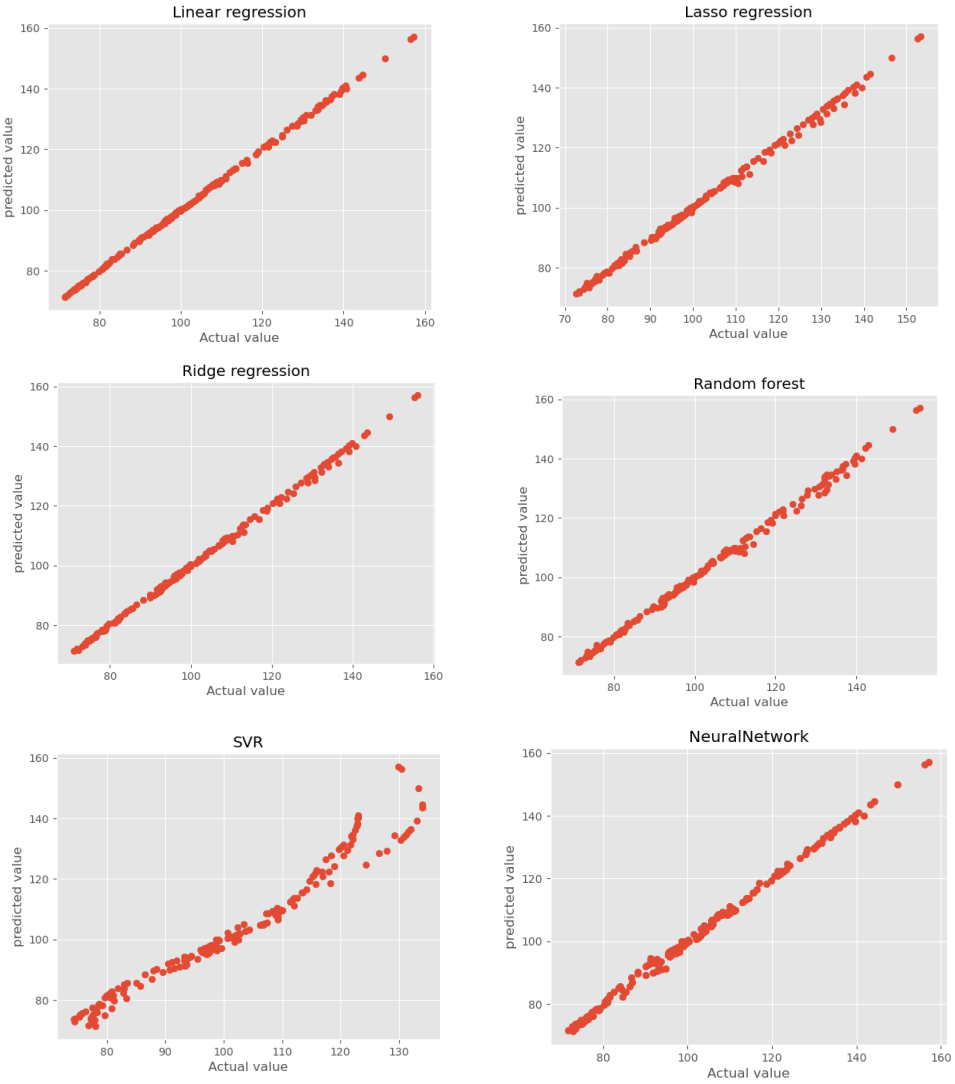


Figure 4.24: Predicted vs. Actual plot for different regression model and Neural network for organic acid prediction with VEAS lab dataset.

4.3.5. Model evaluation:

Below table has shown summary of model performance evaluation for all models for prediction of influent Organic acid with VEAS lab-dataset

Table 4.5: Evaluation of different model with VEAS Lab - dataset

Predicted parameters	**Input parameters**	**Model name**	**R^2**	**mae**	**RMSE**
Organic acid	Flow rate, NH_4-N, pH, 'TS%	SARIMAX	0.95		3.40
		Linear Regression	0.999725	0.263953	0.324187
		Lasso	0.995785	0.980631	1.269679
		Ridge	0.998946	0.507910	0.634970
		Bayesian Ridge	0.999725	0.263864	0.324080
		Decision Tree Regressor	0.996074	0.813642	1.225406
		Random Forest Regressor	0.997599	0.614842	0.958192
		XGB Regressor	0.997588	0.649894	0.960487
		Ada Boost Regressor	0.995674	1.028502	1.286253
		SVR	0.926766	2.930638	5.292422
		Dummy Regressor	-0.006383	15.448105	19.619184
		Neural Network	0.997377	0.778148	1.001653

5 Discussions

To find an effective and efficient way to predict important influent compositions of wastewater and sludge treatment process, an ARIMA model, SARIMAX model, Regression model and an neural network model were developed in this study.

5.1. ARIMA

Despite the few drawbacks that ARIMA possesses, ARIMA has proven to be a reliable model for hydrological purposes, such as flow forecasting. Like all models, the ARIMA model has advantages as well as disadvantages.

Some advantages that ARIMA possess include the fact that it is a simplistic model which can be interpreted and calculated easily. To continue, while most data-driven models typically use information from exogenous variables, no weather data is needed for ARIMA to make its prediction. This is beneficial as not all historical data is accurate or even exists, especially higher frequency data. Also, real-time industrial data usually does not have good correlations with other parameters. For instance, both Lab-analyzed and online dataset has correlation matrix with very low correlation between Influent compositions such as Flow rate PO_4, sCOD, NH_4-N and Flow rate, sCOD, NO_3. ARIMA avoids these issues as it only requires its own historical data of output to forecast output. In this study, ARIMA is successfully applied to the forecasting of inlet PO_4 in HIAS laboratory analyzed dataset with R^2 as 0.72.

Although the ARIMA model has many advantages, it also has some drawbacks. To begin with, ARIMA does not use information from exogenous variables, which leads to a limit of predictability. Meanwhile, ARIMA can only run a continuous time series, meaning the missing values in the dataset must be either filled with interpolation method or need to be dropped. This makes the process long and interpolated data can decrease the model performance as the variations in the data may be incorrectly represented. Moreover, one of ARIMA's assumptions is that there is no seasonality presence in the data. The model accuracy may be affected when this assumption is not satisfied, which is very usual in influent forecasting problems. Another problem of ARIMA model is that it cannot handle the non-linear nature in data and shows very poor prediction result in this case. In our research, ARIMA model showed very poor prediction in influent SCOD prediction with HIAS online dataset. There was lots of noise in the original dataset which was not

removed after removal of outliers. In addition, also after differentiating twice there was still some seasonality or non-linearity present in the data. This makes the ARIMA prediction result not satisfactory with high RMSE score as 11.5 and poor R^2 score of 0.21.(Boyd et al., 2019)

5.2. SARIMAX

The SARIMAX method was applied due to its capability to handle shortcoming of the ARIMA model in dealing with seasonal components in the time series. There has not been many uses of SARIMAX model application in wastewater influent prediction before except inlet and outlet TP prediction by (Ly et al., 2022) . Besides, low temporal resolution forecast of 10 min or 1hour for wastewater or sludge influent with ARIMA family models has not been demonstrated in the literature.

SARIMAX technique was partially successful in wastewater influent modelling and forecasting for the case study WWTP at low temporal resolution with hourly time series data from VEAS lab dataset. SARIMAX (1,1,0)(2,0,0)24 was identified as the best model amongst potential ones. The orders (p,d,q) and (P,D,Q) of the proposed SARIMAX model were diagnostically checked by performing visualization (ACF and PACF graphs), and statistical test (Ljung-Box test) for the residuals. The results indicate the proposed SARIMAX model provides high accuracy forecasts based on several evaluation criteria including RMSE = 3.40 and R^2 = 0.95. The sludge influent forecasts for low temporal resolution of 60 min generated from the proposed SARIMAX model can be utilized as an input for sludge treatment operations optimization model or controller in real-time. Sludge influent composition predictions are an important factor in optimizing the treatment operations. As sludge influent composition is corrosive and it is difficult to use direct sensor for high maintenance, it is quite helpful and cost saving if influent composition concentration can be predicted in advance.

An advantage of the SARIMAX technique is that in addition with the simplicity of SARIMA model methodology, this can include other attributes that have influences on the wastewater and sludge influent composition as its inputs.

However, it is also a limitation of SARIMAX model that if exogeneous inputs have poor correlation coefficient with output, then the model performance decreases which we can see in results of

PO_4 and sCOD prediction with HIAS lab-analyzed dataset. Performance evaluation of SARIMAX with both cases showed poor result with R^2 = 0.43. However, a comparative study on forecasting wastewater and sludge influent composition using SARIMAX model have been done in this paper.(Do et al., 2022)

5.3. Regression methods and Neural network

This study aimed to develop a practical method to predict wastewater and sludge influent composition, which can operate on the individual plant to analyze and identify the problem in real-time.(Rahmat et al., 2022)

This work uses the daily average value of flow rate, PO_4, sCOD, and NH_4-N from the Hias lab dataset, an hourly dataset of COD, flowrate, NO_3 from the online dataset, and an hourly dataset of organic acid, flow rate, NH_4-N, pH, TS% from VEAS lab-dataset to calibrate the regression models. The regression algorithms have demonstrated a reasonably accurate estimation of selected output parameters in all cases. The variations in result observed in some algorithms (Lasso regression or dummy regression) is due to weakness of those models to capture and learn the trend of non-linear dataset.

A possible solution to improve the accuracy of different regression methods is to increase the time resolution of the lab-measured data used in calibrating the MLR models. In this instance, performance evaluation of HIAS online data and VEAS lab data with hourly time resolution showed way better R^2 as 0.88 or 0.99 than HIAS daily lab data with R^2 as 0.77. The MLR models calibrated using higher time resolution can better capture data variations and subsequently improve the estimation results.(Nair et al., 2022)

Data is the most crucial factor of machine learning, and the size of the data set has a significant impact on the accuracy of prediction. In the wastewater and sludge treatment process, the online monitoring system is adopted to obtain the data of influent quality and quantity, which is relatively easy to obtain a large number of training data and one of the main reasons for adopting the machine learning method to predict the influent quality. However, due to the high cost of the analyzer and the cost of application and development of online monitoring systems in wastewater and sludge treatment, the amount of online data is limited, or some parameters are missing from online data. This leads to proceeding with daily average lab analyzed data of small periods (1-

month data for VEAS process), which indirectly leads to low prediction accuracy. This limitation was overcome by including previous steps of all inputs and output in training the model to increase the dataset size, which increases the model's accuracy. For example, accuracy result of MLR has increased from R^2 is 0.22 to R^2= 0.89 with this method.(Wang et al., 2021)

Neural Network

Results of observations and predictions obtained by the multiple-layer ANN model are plotted result section. It is clearly shown that the NN has good generalization ability and could capture the trend of different datasets. The scatter plot of the NN model in the result section shows that most points are close to the diagonal line, indicating that the NN model could be used to predict the influent parameters wastewater and sludge treatment process.

Meanwhile, the NN model has no requirement regarding data stationarity, and exogenous variables can be included for training and prediction purposes. It is worth mentioning that, in this case study, since the range of the training set is larger than that of the testing set, the model's extrapolation ability was not tested. Considering that predictions made beyond the range of the training set tend to be unreliable if the testing data contain values that exceed the training data range in future applications, the model's extrapolation ability should be evaluated to provide a more reliable and reasonable prediction.(Zhang et al., 2019)

6. Conclusion and Future work

This study aimed to develop an efficient and practical method to predict wastewater and sludge influent composition, which any individual plant can incorporate to analyze and estimate the influent quality in real time. This is important for real time control strategy in wastewater and sludge treatment processes.

For this purpose, different datasets from laboratory analysis and online sensor measurements of different time sampling frequencies (10 min, 1 hour, daily) have been collected from two wastewater treatment facilities in Norway, namely the HIAS and VEAS processes.

The characterization and parameters of the three datasets used differed from each other, with a different correlation matrix. In this work, two mathematical model algorithms are used, ARIMA and SARIMAX, the most popular time series analysis in wastewater influent prediction. Compared to previous research, this study only shows a significantly better result for the SARIMAX model with VEAS lab data with high accuracy of R^2 as 0.95.

Besides this, a list of multiple linear regression algorithms (Linear regression, Lasso, Ridge), shallow machine learning algorithms like Decision Tree, Random Forest, SVR, and deep learning algorithms like Multiple layer neural networks were developed to detect and predict parameters such as PO_4 and sCOD from wastewater influent and organic acid from sludge influent.

Despite a significant variation in datasets induced by external activities (rain, snowmelt, human activities, slaughterhouse, and industrial activities), internal activity (error or failure in plant operation), and complicated wastewater and sludge treatment processes, SARIMAX, MLR, and Neural network demonstrated entirely satisfactory performances for model evaluation estimation as evidenced by low values of RMSE and near to one value of R^2. In terms of efficiency, SARIMAX exhibited acceptable results in VEAS dataset, whereas MLR and NN models exhibited overall acceptable result in all three datasets, confirming that the application of these algorithms for influent prediction modeling was successful.

In contrast, ARIMA and SARIMAX algorithms in Hias datasets did not meet the requirements because of the complex and nonlinear structure of the dataset issue.

This work offers a comprehensive comparison between different statistical and machine learning algorithms, it illustrates control strategies to optimize wastewater and sludge treatment efficiency

by predicting the inlet compositions. This study is essential and beneficial when considering the environmental factors of managing wastewater and sludge from households and industries. Despite the impressive results reported here, many limitations should be addressed in future work to successfully apply these algorithms for influent prediction.

It is necessary to collect more data with small time resolution from different WWTPs to ensure the algorithms are generalized.

Deep machine learning architectures such as ANFIS, and LSTM can overcome the issues of addressing large datasets with nonlinear and nonstationary behavior and helping enhance forecasting capability.

If there are ways to enhance the performance of these algorithms, particularly SARIMAX, using additional tools would be beneficial.

Finally, this study of influent prediction plays a vital role and can be further used for controlling the wastewater and sludge treatment process with a control strategy such as stochastic MPC. Stochastic Model Predictive Control (MPC) incorporates stochastic or random variations in the system model. It considers the probability distribution of uncertain parameters in the model, which makes it useful in wastewater treatment processes, where there are often uncertainties in influent flow rates, concentrations, and compositions.

Accurate influent prediction enables the control strategy to adjust the control parameters, optimize the process, minimize the operating costs, improve the effluent quality, and plan maintenance activities effectively. This research plays a role in achieving the zero-pollution goal outlined in the European Green Deal.

References

AlindGupta. (n.d.). *Regularization in Machine Learning*.

Andreides, M., Dolejš, P., & Bartáček, J. (2022). The prediction of WWTP influent characteristics: Good practices and challenges. *Journal of Water Process Engineering*, *49*(July). https://doi.org/10.1016/j.jwpe.2022.103009

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q., & Zhou, P. (2019). Influent forecasting for wastewater treatment plants in North America. *Sustainability (Switzerland)*, *11*(6), 1–14. https://doi.org/10.3390/su11061764

Brendan Artley. (2022, April 26). *Time Series Forecasting with ARIMA , SARIMA and SARIMAX*. Towards Data Science.

Do, P., Chow, C. W. K., Rameezdeen, R., & Gorjian, N. (2022). Wastewater inflow time series forecasting at low temporal resolution using SARIMA model: a case study in South Australia. *Environmental Science and Pollution Research*, *29*(47), 70984–70999. https://doi.org/10.1007/s11356-022-20777-y

DR NILIMESH HALDER. (2019, November 18). *Evaluate Machine Learning Algorithm in R – kfold cross validation in R*.

El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering*, *44*. https://doi.org/10.1016/j.jwpe.2021.102380

Jonassen, K. R., Nåvik, K., Veas, H., Nu, K., Christoffersen, B., & Hult, F. (n.d.). *Arbeidsgruppa Rune Holmstad VEAS ruho@veas.nu*.

Ly, Q. V., Truong, V. H., Ji, B., Nguyen, X. C., Cho, K. H., Ngo, H. H., & Zhang, Z. (2022). Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. *Science of the Total Environment*, *832*(January). https://doi.org/10.1016/j.scitotenv.2022.154930

Nair, A., Hykkerud, A., & Ratnaweera, H. (2022). Estimating Phosphorus and COD Concentrations Using a Hybrid Soft Sensor: A Case Study in a Norwegian Municipal Wastewater Treatment Plant. *Water (Switzerland)*, *14*(3). https://doi.org/10.3390/w14030332

Rahmat, S., Altowayti, W. A. H., Othman, N., Asharuddin, S. M., Saeed, F., Basurra, S., Eisa, T. A. E., & Shahir, S. (2022). Prediction of Wastewater Treatment Plant Performance Using Multivariate Statistical Analysis: A Case Study of a Regional Sewage Treatment Plant in Melaka, Malaysia. *Water (Switzerland)*, *14*(20). https://doi.org/10.3390/w14203297

Rob J Hyndman and George Athanasopoulos. (n.d.). *Forecasting: Principles and Practice* (3rd edition).

Selva Prabhakaran. (2019, November 2). *Augmented Dickey Fuller Test (ADF Test) – Must Read Guide*.

Veena Ghorakavi. (n.d.-a). *Types of Regression Techniques in ML*.

Veena Ghorakavi. (n.d.-b). *Types of Regression Techniques in ML*.

Wang, R., Pan, Z., Chen, Y., Tan, Z., & Zhang, J. (2021). Influent quality and quantity prediction in wastewater treatment plant: Model construction and evaluation. *Polish Journal of Environmental Studies*, *30*(5), 4267–4276. https://doi.org/10.15244/pjoes/132821

Wodecka, B., Drewnowski, J., Białek, A., Łazuka, E., & Szulżyk-Cieplak, J. (2022). Prediction of Wastewater Quality at a Wastewater Treatment Plant Inlet Using a System Based on Machine Learning Methods. *Processes*, *10*(1). https://doi.org/10.3390/pr10010085

Zhang, Q., Li, Z., Snowling, S., Siam, A., & El-Dakhakhni, W. (2019). Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. *Water Science and Technology*, *80*(2), 243–253. https://doi.org/10.2166/wst.2019.263

Cheng, T., Harrou, F., Kadri, F., Sun, Y., & Leiknes, T. (2020). Forecasting of wastewater treatment plant key features using deep learning-based models: A case study. IEEE Access, 8, 184475–184485. https://doi.org/10.1109/ACCESS.2020.3030820

Wang, X., Kvaal, K., & Ratnaweera, H. (2019). Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. Journal of Process Control, 77, 1–6. https://doi.org/10.1016/j.jprocont.2019.03.005

List of Attachments

- Appendix
- ARIMA MODEL with HIAS lab-dataset- PO4 prediction.ipynb
- ARIMA - HIAS online data - sCOD prediction.ipynb
- SARIMAX-Hias labdata - PO4 prediction.ipynb
- SARIMAX-Hias labdata -sCOD prediction.ipynb
- SARIMAX-Veas labdata -organic acid prediction.ipynb
- Regression and NN - Hias labdata - PO4 prediction.ipynb
- Regression and NN -VEAS labdata - Organic acid prediction.ipynb
- Regression and NN with online data -HIAS-SCOD prediction.ipynb

Appendix

Table A.1 : Taxonomy on literature review

Predicted parameters(Quantity and Quality prediction)	Input parameters	Mathematical approach/ Model used	Accuracy	Literature reference
Quantity: Flow rate Quality: COD and NH3-N	precipitation, temperature (Exogeneous input) influent flow, COD, and ammonia	Linear Regression, Ridge Regression, ElasticNet Regression Lasso Regression	Quantity prediction: 86.19% COD conc. pred.:82% NH3-N conc. pred: 82%	10.15244/pjoes/132821
Outlet TP(Total phosphate)	Influent water flow(IWF), pump level(PPL),Inlet-COD, Inlet-NH3, Inlet-TP, Inlet-TN ,Outlet-COD ,Outlet-NH3 ,Outlet-TP ,Outlet-TN	SARIMAX, Random Forest (RF), Support Vector Machine (SVM), Gradient Tree Boosting (GTB), Adaptive Neuro-Fuzzy Inference System (ANFIS) Long Short-Term Memory (LSTM)	R^2= 92%	http://dx.doi.org/10.1016/j.scitotenv.2022.154930
Flow rate, COD, SS, TN and TP	COD, TN and TP	k-Nearest neighbor	MAPE <8.9 % for all parameters on both wet and dry days	10.1007/s11783-015-0825-7
Total Phosphorus (TP) and Chemical Oxygen Demand (COD) in both influent and effluent	pH, TSS, Flowrate, conductivity, ORP, PAX and Polymer dose flow-meters	Artificial Neural Network (ANN), Support vector machine(SVM), Enesemble Tree(ET), Multiple Linear Regression(MLR) Nelder–Mead (NM) . Trust-region Newton conjugate gradient (TR) Sequential Least Squares Quadratic Programming (SLSQP) and Broyden–Fletcher–Goldfarb–Shanno (BFGS)	TP: SLSQP: R^2= 76% COD: SLSQP: R^2= 70%	https://doi.org/10.3390/w14030332

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	702
Model:	SARIMAX(1, 1, 4)x(1, 0, [1], 7)	**Log Likelihood**	-3880.700
Date:	Fri, 28 Apr 2023	**AIC**	7781.401
Time:	17:01:33	**BIC**	7826.926
Sample:	03-01-2021	**HQIC**	7798.998
	- 01-31-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
NH4 inn	3.7245	0.192	19.379	0.000	3.348	4.101
PO4 Inn Linje 1	32.9623	2.102	15.684	0.000	28.843	37.082
ar.L1	-0.9462	0.064	-14.842	0.000	-1.071	-0.821
ma.L1	0.6116	0.069	8.862	0.000	0.476	0.747
ma.L2	-0.5939	0.041	-14.466	0.000	-0.674	-0.513
ma.L3	-0.5516	0.038	-14.352	0.000	-0.627	-0.476
ma.L4	-0.2952	0.034	-8.784	0.000	-0.361	-0.229
ar.S.L7	0.9719	0.012	77.906	0.000	0.947	0.996
ma.S.L7	-0.8962	0.026	-34.079	0.000	-0.948	-0.845
sigma2	3363.6880	106.169	31.682	0.000	3155.601	3571.775

Ljung-Box (L1) (Q):	1.29	**Jarque-Bera (JB):**	353.93
Prob(Q):	0.26	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.47	**Skew:**	0.13
Prob(H) (two-sided):	0.00	**Kurtosis:**	6.47

SARIMAX Results

Dep. Variable:	SCOD Inn	No. Observations:	561
Model:	SARIMAX(1, 1, 4)x(1, 0, [1], 7)	Log Likelihood	-3106.359
Date:	Fri, 28 Apr 2023	AIC	6232.718
Time:	17:03:19	BIC	6275.997
Sample:	03-01-2021	HQIC	6249.618
	- 09-12-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
NH4 inn	3.5613	0.285	12.477	0.000	3.002	4.121
PO4 Inn Linje 1	31.4866	2.569	12.255	0.000	26.451	36.522
ar.L1	-0.9364	0.129	-7.232	0.000	-1.190	-0.683
ma.L1	0.5918	0.161	3.679	0.000	0.277	0.907
ma.L2	-0.6431	0.150	-4.278	0.000	-0.938	-0.348
ma.L3	-0.6257	0.092	-6.799	0.000	-0.806	-0.445
ma.L4	-0.3223	0.055	-5.843	0.000	-0.430	-0.214
ar.S.L7	0.9416	0.026	36.091	0.000	0.891	0.993
ma.S.L7	-0.8005	0.044	-18.227	0.000	-0.887	-0.714
sigma2	3819.2728	358.367	10.657	0.000	3116.887	4521.659

Ljung-Box (L1) (Q):	0.59	Jarque-Bera (JB):	324.59
Prob(Q):	0.44	Prob(JB):	0.00
Heteroskedasticity (H):	1.77	Skew:	-0.05
Prob(H) (two-sided):	0.00	Kurtosis:	6.73

Figure A1:Best fit SARIMAX model for sCOD prediction with HIAS lab data

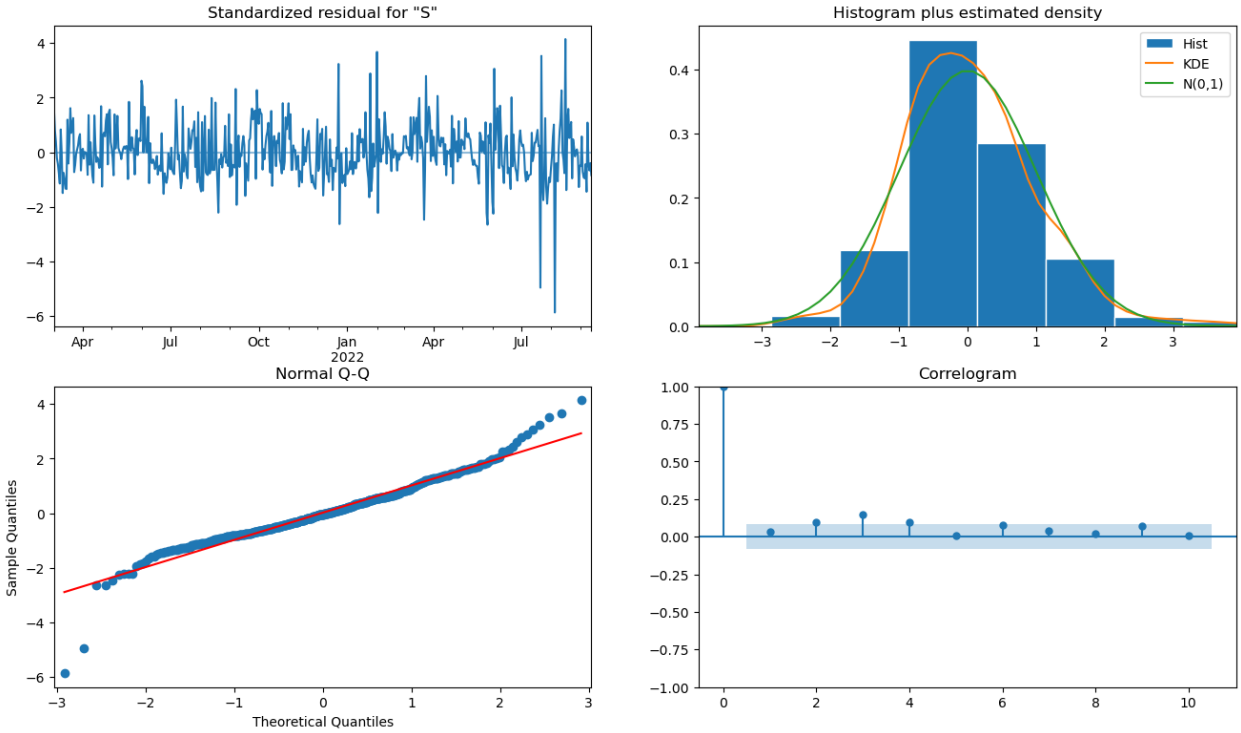


Figure A2: Diagnostic check for SARIMAX model For sCOD prediction with HIAS labdata

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=64877.007, Time=5.71 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=66897.352, Time=0.15 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=65124.049, Time=0.34 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=65541.458, Time=0.99 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=66895.367, Time=0.14 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=65036.196, Time=1.41 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=65077.667, Time=2.40 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=64869.676, Time=5.44 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=64867.751, Time=4.66 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=65042.321, Time=0.55 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=64869.649, Time=7.36 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=65089.460, Time=0.52 sec
 ARIMA(4,1,0)(0,0,0)[0] intercept   : AIC=64979.347, Time=0.90 sec
 ARIMA(4,1,2)(0,0,0)[0] intercept   : AIC=64868.919, Time=10.38 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=64865.757, Time=2.27 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=65075.672, Time=1.20 sec
 ARIMA(3,1,0)(0,0,0)[0]             : AIC=65040.326, Time=0.31 sec
 ARIMA(4,1,1)(0,0,0)[0]             : AIC=64867.655, Time=3.53 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=64867.682, Time=2.77 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=65087.465, Time=0.27 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=64875.012, Time=3.28 sec
 ARIMA(4,1,0)(0,0,0)[0]             : AIC=64977.353, Time=0.46 sec
 ARIMA(4,1,2)(0,0,0)[0]             : AIC=64866.925, Time=5.13 sec

Best model:  ARIMA(3,1,1)(0,0,0)[0]
Total fit time: 60.187 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	8819
Model:	SARIMAX(3, 1, 1)	**Log Likelihood**	-32427.878
Date:	Sat, 13 May 2023	**AIC**	64865.757
Time:	17:39:35	**BIC**	64901.180
Sample:	0	**HQIC**	64877.823
	- 8819		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2678	0.014	91.345	0.000	1.241	1.295
ar.L2	-0.2497	0.007	-38.069	0.000	-0.263	-0.237
ar.L3	-0.1328	0.003	-39.042	0.000	-0.139	-0.126
ma.L1	-0.8857	0.013	-67.523	0.000	-0.911	-0.860
sigma2	91.5543	0.303	302.005	0.000	90.960	92.148

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	5053663.67
Prob(Q):	0.97	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.91	**Skew:**	1.73

Figure A33:ARIMA potential models sCOD prediction with Hias online dataset

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	673
Model:	SARIMAX(1, 1, 0)x(2, 0, 0, 24)	**Log Likelihood**	345.950
Date:	Mon, 24 Apr 2023	**AIC**	-675.900
Time:	16:19:08	**BIC**	-639.818
Sample:	07-01-2022	**HQIC**	-661.926
	- 07-29-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
Flow rate(l/s)	-1.6637	0.296	-5.625	0.000	-2.243	-1.084
NH4-N FOR	0.1692	0.002	78.912	0.000	0.165	0.173
pH FOR	-42.0747	0.996	-42.238	0.000	-44.027	-40.122
TS% FOR	7.7281	0.759	10.186	0.000	6.241	9.215
ar.L1	0.9748	0.024	39.914	0.000	0.927	1.023
ar.S.L24	-0.7878	0.026	-30.440	0.000	-0.839	-0.737
ar.S.L48	-0.3446	0.024	-14.306	0.000	-0.392	-0.297
sigma2	0.0200	0.000	47.484	0.000	0.019	0.021

Ljung-Box (L1) (Q):	0.10	**Jarque-Bera (JB):**	1560892.72
Prob(Q):	0.75	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.53	**Skew:**	-10.66
Prob(H) (two-sided):	0.00	**Kurtosis:**	238.14

Fit the SARIMAX model with the selected order:

SARIMAX Results

Dep. Variable:	Organiske syrer	No. Observations:	538
Model:	SARIMAX(1, 1, 0)x(2, 0, 0, 24)	Log Likelihood	671.075
Date:	Wed, 26 Apr 2023	AIC	-1326.149
Time:	10:12:56	BIC	-1291.861
Sample:	07-01-2022	HQIC	-1312.736
	- 07-23-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
Flow rate(l/s)	-0.7242	0.196	-3.698	0.000	-1.108	-0.340
NH4-N FOR	0.1101	0.001	211.587	0.000	0.109	0.111
pH FOR	-40.5885	0.440	-92.283	0.000	-41.451	-39.726
TS% FOR	-4.3446	0.299	-14.550	0.000	-4.930	-3.759
ar.L1	0.9763	0.017	57.839	0.000	0.943	1.009
ar.S.L24	-0.3048	0.026	-11.855	0.000	-0.355	-0.254
ar.S.L48	-0.0853	0.028	-3.076	0.002	-0.140	-0.031
sigma2	0.0048	0.000	45.974	0.000	0.005	0.005

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	1315155.42
Prob(Q):	0.75	Prob(JB):	0.00
Heteroskedasticity (H):	0.29	Skew:	-13.07
Prob(H) (two-sided):	0.00	Kurtosis:	244.03

Figure A4:Best Model to fit SARIMAX model prediction of Organic acid with VEAS lab-data