Environment and Ecology Research 11(3): 456-466, 2023
DOI: 10.13189/eer.2023.110305 http://www.hrpub.org

Application of ARIMA Imputation Model and RNN Forecasting Model – WWTPs Performances Optimization

A. Chaoui[1,*], W. Elkhoumsi[1], M. Laaouan[2], R. Bourziza[1], K. Sebari[1]

[1]Hassan II Institute for Agronomy and Veterinary Medicine, Rabat, Morocco
[2]International Institute for Water and Sanitation, Rabat, Morocco

Received March 2, 2023; Revised April 28, 2023; Accepted May 19, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] A. Chaoui, W. Elkhoumsi, M. Laaouan, R. Bourziza, K. Sebari , "Application of ARIMA Imputation Model and RNN Forecasting Model – WWTPs Performances Optimization," Environment and Ecology Research, Vol. 11, No. 3, pp. 456 - 466, 2023. DOI: 10.13189/eer.2023.110305.*

(b): *A. Chaoui, W. Elkhoumsi, M. Laaouan, R. Bourziza, K. Sebari (2023). Application of ARIMA Imputation Model and RNN Forecasting Model – WWTPs Performances Optimization. Environment and Ecology Research, 11(3), 456 - 466. DOI: 10.13189/eer.2023.110305.*

Abstract Machine learning forecasting has been widely used in order to increase wastewater treatment plants (WWTPs) performance and provide support to WWTPs management. Forecasting wastewater influent quality parameters is not only beneficial for operators and the plant itself, but it is also important environmentally, economically, and socially. In this paper, Chefchaouen's WWTP BOD5 variable will be used as a case example. The current paper applies ARIMA imputation model to have a complete time series variable and complete the dataset that will enable conducting a machine learning forecasting model that is Recurring Neural Networks (RNN) in order to provide accurate predictions. The aim of this paper is to assess the impact of the application of these models on providing support to the plants management and control. In addition to that, analyses will also assess the impact of implementing these models on the use of energy and the injection of oxygen. The models used are statistically correct. The forecasted BOD5 values were close to the actual provided values. BOD5 predictions were converted in order to suggest the total energy consumption per day as well as the total oxygen to be injected. Energy consumption could have decreased in the period of assessment by a percentage of 37% while the oxygen injected could have decreased by a percentage of 90%. Finally, this paper concludes with a discussion as well as the limitations of this work.

Keywords BOD5, Forecasting, Wastewater, Machine Learning, ARIMA, RNN, Energy Consumption, Oxygen Injection

1. Introduction

This paper introduces an imputation model and a forecasting model that is based on machine learning. These models are ARIMA and Recurring Neural Networks (RNN). Machine learning forecasting has been widely used in order to increase WWTPs performance. In the contribution of Al-Asheh et al. [5], the authors applied different forecasting models that include Artificial Neural Networks (ANN), Adaptive Linear Neuron Networks (ADALINE), and Multi-Layer Feed-forward neural networks (ML-FF) to forecast influent quality parameters (BOD, COD, and TSS). Empirical findings indicate that the ML-FF model resulted in lower error terms. But concerning the contribution of Kriger and Tzoneva [20], authors used the ML-FF and Elman Recurring neural networks in order to forecast wastewater treatment plant influent disturbances.

The current paper will present the application of performing models in the case of the Chefchaouen WWTP. The aim of this contribution is to assess the impact of the

application of these models on providing support to the plants' management and control. In addition to that, analyses will also assess the impact of implementing these models on the use of energy and the injection of oxygen. For this, the following section will introduce the plant, the materials and methods, followed by the results sections. Finally, this paper will conclude with a discussion as well as the limitations of this work.

2. Materials and Methods

It is essential to note that in the case of this contribution, the only used forecasted models are time series. By definition, time series variables refer to variables that are measured over time over successive times with a unified frequency such as daily, monthly, quarterly, annually, and whatnot.

Forecasting models, as summarized in the following [7,8,10-12,14,15,21,24,26,28,31,34]:

- Extrapolative methods with simple moving average, exponential smoothing, such as the Holt-Winters method, and autoregressive moving average (ARMA) - aka Box-Jenkins.
- Explanatory variable methods with regression analysis, predictive modelling, artificial neural networks and econometric modelling.
- Simulation modelling methods with cell-based modelling, system dynamics simulation and multi-agent simulation.
- Judgmental methods
- Composite methods with Bayesian forecasting and others like the combination of different forecasting methods and models a type of composite method.

To begin with, it is important to note that predicting wastewater influent parameters confirms that it is not only beneficial for operators and the plant itself, but it is also important environmentally, economically, and socially [6].

Shahidah et al. [32] used a modification of the classical fuzzy time series by applying weighted subset hood Fuzzy Time Series (WeSuFTS) in the water management industry (water production and energy consumption) to measure the performance of the proposed model, the forecast results of the proposed model were compared with the forecast results of Chen and Cheng models. The WeSuFTS model delivered a good performance in the evaluation part with MAPE 0.9252% and APE value between 0.02226% and 1.7877%.

Adib et al. [4] focus on data split of water level time series datasets in producing excellent forecasts as measured by coefficient correlation and based on the local linear approximation method. The result obtained was in the range of strong forecast using chaos approach with over 95% accuracy in every dataset.

The contributions of Zhou et al. [36,37] indicate that having reliable predictions provides support to wastewater management and planning at various stages of the treatment process. However, the contribution of Kim et al. [17] indicates that there are no well-established forecasting tools available for WWTPs operators and managers, which leads these latter to either use their knowledge to make predictions, or use complex physical models that are known for being hard to monitor.

Reagan [30] was among the first scholar to apply the non-seasonal and seasonal ARIMA models to forecast wastewater treatment process variables such as biochemical oxygen demand (BOD), COD, and total suspended solids (TSS). But in the contribution of Boyd et al. [6], authors used the ARIMA model to forecast wastewater inflow on five plants in North America using a dataset that covers a period of two years with different frequencies for each plant. In this contribution, the resulting models were compared across these stations using the RMSE, MAPE, and coefficient of determination. Empirical findings indicate that error terms held a low bias since the RMSE had values that ranged between 0.015 and 0.559, the MAPE had values between 5.371 and 48.959, and the R-squared had values between 0.593 and 0.775.

For Abunama and Othman [2], they used an ARIMA model selected based on the lowest normalized Bayesian Information Criterion (BIC) to forecast a set of influent parameters such as SS, BOD, COD, and ammoniacal nitrogen (NH3-N). Results indicate that the model is suitable as the linear regression analysis between the predicted and observed values resulted in a coefficient of 0.83. In addition to that, the authors Al-Asheh, Mjalli, and Alfadala [5] also used the ARIMA model and compared it to a Multi-layer Feedforward (ML-FF) neural networks model, and the Adaptive Linear Neuron networks neural networks (ADALINE) model. The empirical findings indicate that the ML-FF predictions were more reliable than the other models.

Concerning the contribution of Kim et al. [16], authors used a dataset of 250 observations to forecast the WWTP flow rate, COD, $NH_4^+ - N$, and $PO_4^{3-} - P$. The same contribution trained 150 data points to propose three ARIMA forecasting methods that are: seven models and one-step ahead forecasting (Method 1), one model and one-step head forecasting (Method 2), and one-step ahead forecasting only without accumulated error (Method 3). The last method resulted in predictions with low error terms compared to other methods.

In WWTPs, forecasting models were also applied in real-time to forecast the NH_4^+ concerntration using historical data and data collected using the ion-selective sensors [16]. This paper proposed an ensemble-based model that is based on Fourier series to predict the ammonium concentration in sewer systems in order to increase the plant capacity on wet days. The performance of this model provided good simulations of the ammonium loads at the beginning of the rain period, and at the dilution induced by the wet-weather.

Kim et al. [17], applied the k-nearest neighbor

forecasting method in order to forecast the WWTP influent characteristics in dry and wet weather conditions. These characteristics account for COD, SS, nitrogen, and phosphorus, and suspended solids. Results indicate that the difference in predictive accuracy in dry weather was less than 5%, and larger for wet weather. Yet, in this contribution, there is enough evidence to conclude that the model resulted in good predictions that can provide support to plant management and control.

With regards to the contribution of Kim et al. [19], the artificial neural network (ANN) and the $ASM3 + Bio - P$ models were used in order to forecast the influent and effluent in order to forecast the $NH_4^+ - N$. Furthermore, this contribution forecasted the control of the effluent water quality in the A^2/O (Anaerobic/Anoxic/Oxic) process 1 day in advance. Empirical findings indicate that by applying this model, the $NH_4^+ - N$ concentration decreased by 68% by only increasing the air flow rate by 9%, as suggested by the model.

For Nadiri, Shokri, Tsai, and Moghaddam [27], the application of forecasting models consisted of applying a proposed ensemble supervised committee fuzzy logic model to predict various wastewater influent parameters that are BOD, COD, and TSS. The used model is based on artificial neural networks and uses the MAPE as an error comparison tool. Results indicate that the proposed model performs better than individual fuzzy logic models.

With regards to the publication of Yu et al. [35], it has used a combined Kernel Principal Component Analysis (KPCA) and Extreme Machine Learning (EML) to forecast the inlet water quality of sewage treatment. This same contribution was compared to a back propagation neural network (BPNN) model and many other models. These models were compared based on the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE). Findings indicate that forecasts resulting from the KPCA model were more accurate than any other model included in the study.

Concerning the contribution of Lotfi et al. [25], many wastewater parameters were forecasted and included BOD, COD, TSS, and total dissolved solids (TDS). This contribution suggested a new methodology that consists of a combination of ARIMA, which is a stochastic model, and a nonlinear outlier robust extreme machine learning technique (ORELM). More than 144 × 8 linear models were considered (using ARIMA) in addition to more than 48 nonlinear models (using ORELM) and 48 hybrid models (using ARIMA-ORELM). Results show that the different obtained correlations between the observed values and the predicted ones could reach a value of 0.99.

For Abba, Nourani, and Elkiran [1], different variables such as PH and SS were forecasted using different models. These latter include general regression neural network (GRNN), Hammerstein-wiener (HW) and non-linear autoregressive with exogenous (NARX) neural network, and the least square support vector machine (LSSVM), besides many other nonlinear ensemble techniques. Findings indicate that each model was suitable for a specific variable. For instance, HW model showed better accuracy in forecasting SS while GRNN-E model was better in predicting pH and SS.

The contribution of existing literature is still large. Still, it is important that there are different models that could potentially be used in the case of this dissertation. These models will be presented in a later section.

2.1. Chefchaouen WWTP Description

The Chefchaouen, Morocco, WWTP (Figure 1) is designed for a capacity of 10,000 m3 per day and allows the treatment of suspended solids, chemical oxygen demand, biological oxygen demand. This is throughout a pre-treatment stage and a biological treatment of the activated sludge type.

Figure 1. Chefchaouen WWTP

The raw wastewater flow rates used are as follows (Table 1):

Table 1. Raw wastewater flow rates

Parameters	Units	Values
Nature of raw water	Domestic	-
Population	Eq./Inhabitant	50 000
Average daily flow	m3/j	10 000
BOD5	Kg/j	1500
TSS	Kg/j	1250
COD	Kg/j	3600

The average daily concentrations of raw wastewater are as follows (Table 2):

Table 2. Average daily concentrations of raw wastewater

Parameter	Unit	Value
BOD5	mg/l	150
COD	mg/l	360
SS	mg/l	125

The following section will introduce all the treatment processes included in this wastewater treatment plan. This is divided into water and sludge sectors.

The water queue is made up of two treatment queues, one of which is backup (in the case of pretreatment), to

give the station flexibility and security in maintenance operations. It consists of the following stages:

- Pretreatment: The pretreatment structures are designed for the reception of peak flow in rainy weather, that is to say, 835m 3 / h:
  - Entrance chamber: raw water is received in this chamber, where there is a weir for the station bypass, as well as the entry of the supernatant and drain pipes. From this chamber, distribution is made at the entrance to the coarse screening channels.
  - Coarse screening: this stage is ensured by two channels, one of which is emergency. Each channel is provided with an inclined screen with automatic cleaning of a 50mm air gap and isolated by wall valves at the inlet and outlet. Then, a screw transports the waste from the grids to a conveyor compactor which sends it to a bucket for disposal.
  - Lifting station: the transfer of wastewater to biological treatment is ensured by a pumping station composed of four submersible pumps, one of which is installed backup unit flow of 280 m3 / h. The water is led to the fine screen through a DN 400 stainless steel discharge pipe fitted with equipment for flow measurement (electromagnetic flowmeter)
  - Fine screening: this stage is made up of two channels, one of which is emergency. Each channel is provided with an inclined screen with automatic cleaning of a 20mm air gap and insulated by cofferdams at the entrance and exit. The waste from the grids passes through a conveyor screw and thereafter will be sent to a bucket for their evacuation by means of a conveyor compactor.
  - Desanding-oiling: consisting of two structures in the form of an aerated type longitudinal channel with a tranquilization zone for the separation of grease and collection of sand. Aeration is done by submerged aeroflot aerators, two per structure. The sand is extracted by a centrifugal pump to a sand classifier with a capacity of 30m3 / h for its treatment. The fat is sent to a fat concentrator.
  - Distribution to biological reactors: After the grit separators, the water is sent to the biological reactors by means of a distribution structure.
- Biological treatment: Biological treatment is provided by two biological reactors. These reactors are of rectangular type with a unit volume of 2287m3, with two zones: anoxic zone and an oxic zone. The aeration of the oxic zones is done through diffusers with fine bubbles. The anoxic zone is equipped with submersible agitators in order to keep the biomass in suspension. Sludge recirculation is ensured by two submersible pumps per line, one of which is a back-up unit with a flow rate of 315m3 / h.
- Anoxic Zone: The purpose of this zone is to optimize the mixing of raw water with recirculated sludge, in order to promote the development of flock bacteria to the detriment of filamentous bacteria. Indeed, the book dedicated to the contact area is strongly recommended. To avoid the formation of filamentous bacteria, the raw water and part of the recirculated sludge are brought into contact by mechanical agitation.
- Degassing: At the outlet of the reactors, the water passes through two degassing structures where fast submersible agitators are installed. The two structures are connected to give the station even more flexibility. These works are intended for:
  - Limiting the formation of air plugs in the transit pipe to the clarifiers.
  - Significantly reducing foaming on clarifiers and contributing to improving the quality of purified water by limiting the entrainment of SS towards the surface of the clarifiers.
- Division: After degassing, a distribution structure is produced to ensure the same input flow to the secondary settlers.
- Secondary settling: Secondary decantation is ensured by two secondary decanters. These decanters are of circular type with a unit diameter of 28.30 m. The bottom scraping is ensured by a scraper bridge.
- Flow measurement: The flow measurement of treated water is done using an electromagnetic flow meter DN 500.

The sludge queue is made up of tree treatment queues. It consists of the following stages:

- Extraction of excess sludge: The excess sludge is pumped by two pumps, one of which is an emergency backup installed of 10m3/h per line, towards the thickener.
- Thickening of sludge: This operation is ensured by a dynamic thickener with gravity concentration of the sludge and motorized central drive with a diameter equal to 7.15m.
- Mechanical sludge dewatering: Dehydration is ensured by two centrifuges with a capacity of 6.7m3 / h. These centrifuges are powered by two screw pumps, one of which is emergency and has a unit flow rate of 8m3 / h. The sludge is conditioned by the preparation of the polymer in an automatic preparation center powered by two screw pumps, one of which is emergency and has a unit flow rate of 700 l/h. The dehydrated sludge is sent by a conveyor screw, where mixing takes place with the slaked lime, to a force-feeding pump which leads them to a storage silo with a volume equal to 25m3.

2.2. Data

The data used in this contribution is from Chefchaouen WWTP between 09/05/ 2017 and 28/02/2018. The data set accounts for the wastewater influent and effluent parameters that are:

- BOD5 influent and effluent quality variable: The values of this variable are stated in mg/L and will enable calculating the daily treated BOD5 variable.
- Wastewater inflow and outflow: These variables will enable converting the BOD5 variables unit from mg/L to kg per day.
- Daily energy consumption: This variable will be used to assess the impact of implementing the forecasting model on energy consumption.
- Daily oxygen injection: This variable will be used to assess the impact of implementing the forecasting model on the oxygen injection.

2.3. Methods

The dataset accounts for time series variables that have missing values. For this, the first part of this contribution will apply the ARIMA data imputation model. Concerning the second part of this contribution, it will predict the BOD5 values using the Recurring Neural Network model. This model will use an iterative process to forecast the values of the next 5 days after the training and testing phase. In order to predict the value of the 6th day, the model assumes that the laboratory of the plant has the experimental BOD5 value of the day after the testing and training period. For this, the model will replace the forecasted value of (y1) and replace it with the actual BOD5 value of this corresponding day. This iterative process will be used for the remaining forecasted periods as illustrated in Figure 2.

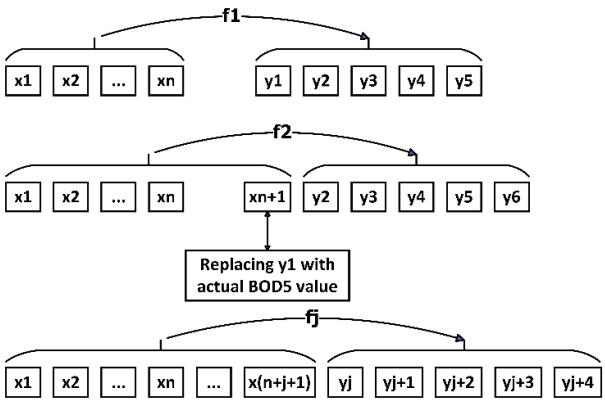

Figure 2. BOD5 iterative forecasting process

Where:

- xi: represents the actual BOD5 variables.
- yj: represents the forecasted BOD5 variables.
- fj; represents each RNN model to predict the periods between yj and yj+4.

2.4. Theoretical Model

The following contribution represents a case study based on the Chefchaouen plant. This contribution will apply ARIMA data imputation model to replace missing values and the Recurrent Neural Network to generate forecasts.

As was mentioned in the data section, the period of the time series provided data will be divided into a train and a test period, while the second period will be used to assess the impact of applying these models on the actual plant's energy and oxygen injected performance.

Figure 3 summarizes the theoretical framework that will be used in this contribution. The first and second steps consist of applying the ARIMA data imputation model and forecasting using the RNN model. These steps will enable forecasting BOD5 values that will be converted using wastewater inflow, wastewater outflow, and BOD5 effluent quality variable into the treated BOD5 in terms of kg per day. This latter variable will be used to suggest daily energy consumption and oxygen injection. In order to assess the difference between the mean of the actual energy consumption, oxygen injections and their corresponding suggested values, the t-statistic test will be used.

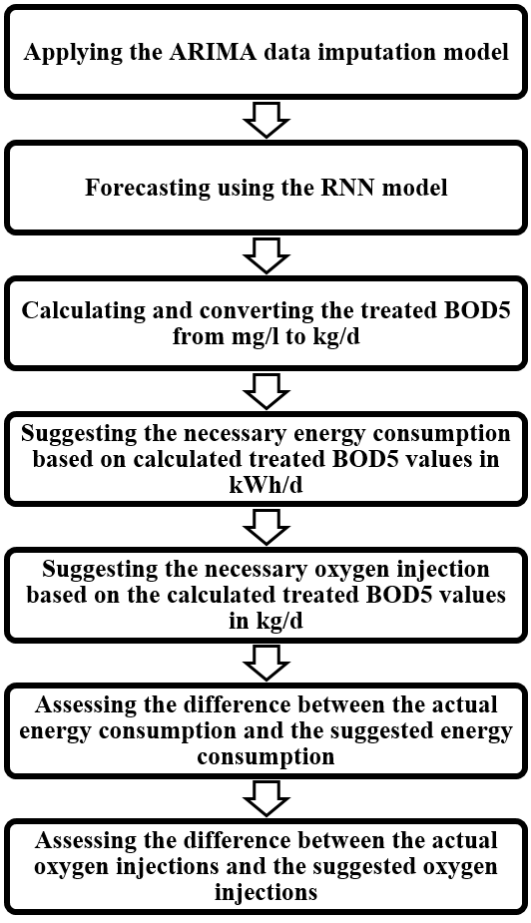

Figure 3. Theoretical framework to assess the impact of forecasting BOD5 values on energy consumption and oxygen consumption

For each comparison made, the null and alternative

hypotheses are stated as [18]:

- H0: $\mu_a = \mu_b$, the means of the two samples are equal.
- Ha: $\mu_a \neq \mu_b$, the means of the two samples are different.

In order to test for this difference, the t-test statistic assumes that the variance between the original values and the suggested ones is different, thus, the test can be mathematically expressed such as [18]:

$$t = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \quad (1)$$

Where:

- $\overline{x_a}$: is the mean of the first sample.
- $\overline{x_b}$: is the mean of the second sample.
- s_a^2: is the variance of the first sample.
- s_b^2: is the variance of the second sample.
- n_a: is the sample size of the first sample.
- n_b: is the sample size of the second sample.

2.4. Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a univariate model that is used to predict future behaviors based on the historical observations of the variable itself. This model is the combination of the autoregressive model (AR) and the moving average model (MA), and is often referred to by scholars as ARIMA(p,d,q), where p refers to the AR order, q refers to the order of the MA model, and d refers to the degree of differencing. In the contribution of Adhikari & Agrawal [3], authors indicate the general formula of the ARIMA model that is given such as:

$$\varphi(B)z_t = \emptyset(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \quad (2)$$

Where:

- θ_0: is a constant;
- z_t: discrete time series value at t;
- a_t: white noise process;
- B: backward shift operator ($Bz_t = z_{t-1}$);
- $\varphi(B)$: nonstationary autoregressive operator;
- $\theta(B)$: moving average operator.

But in the contribution of Shakti et al. [33], authors define the different stages of the ARIMA process in different steps that are summarized in Figure 4. The first step consists of checking the stationarity of the time series variable that will be used. This is in order to satisfy the main assumption of the ARIMA model that will be represented in a later section. In the case where the data does not meet this assumption, the use of the differencing method or data transformation is required. The next step, consists of plotting the Auto-Correlation Function (ACF) and the Partial Auto-Correlation Function (PACF) in order to determine the different parameters of the AR and MA models. This is followed by the estimation of the integration level. However, in order to find the ARIMA model of the best fit, it is very common to generate a simulation that trains different ARIMA models with different parameters in order to select the best one based on the best residual analysis. Thus, and after selecting the best model, predictions can be generated.

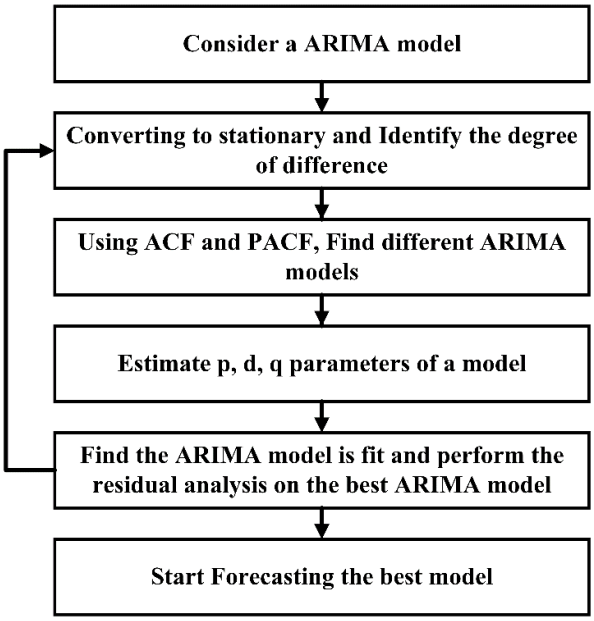

Figure 4. Flow diagram of ARIMA model

In addition to that, the diagram above can be summarized in three steps that consist of model identification, model estimation, and diagnostic checking.

2.5. Recurring Neural Networks

This section introduces the RNN model that is used for sequential data. The contribution of Hochreiter and Schmidhuber [13] indicate that this type of models uses a long short-term memory that is considered to be a gated unit for neural networks. RNN model has 3 gates that manage the memory's content. These gates are given by simple logistics functions of weighted sums that are learnt by back propagation [29]. This is illustrated in Figure 5 [23].

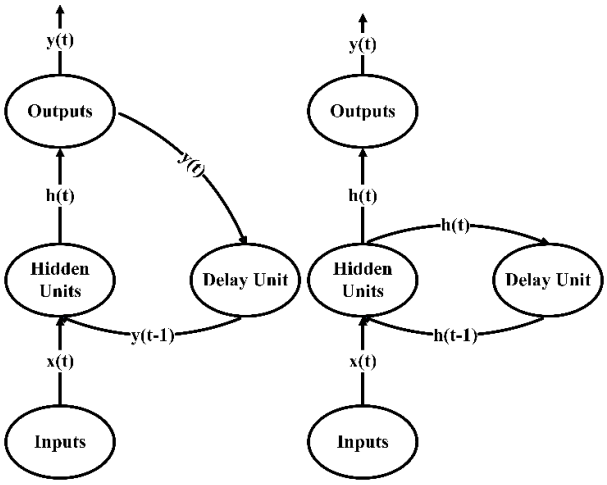

Figure 5. Two simple recurrent network structures

In this process, the cell state is managed by the input gate and the forget gate (or the long-term memory). In addition to that, the input state results in a hidden state. This can be mathematically expressed such as [29]:

- Input gate:

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \quad (3)$$

- Forget gate:

$$f_t = sigmoid\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (4)$$

- Cell state:

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

- Output gate:

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$

- Hidden state:

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

RNN models can learn as well as memorize the time series characteristics. However, it is important to encode many features such as trend and seasonality. For this, it is necessary to normalize the values of the dataset such as [29]:

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (8)$$

Or

$$\tilde{x} = \frac{x - \text{mean}(x)}{sd(x)} \quad (9)$$

The main features of RNN models are lags, trend, seasonality, and dummy indicators. Concerning the lag attribute, they are used as the main input for forecasting. Yet, the only drawback in using this feature is losing the first observation. Still, the higher is the number of the observations, the lower the impact of this feature is. But with regards to the trend attribute, it refers to the ability of the model to learn the trend of the time series over time and duplicate it. This is also the case of the seasonal attribute. But for the dummy variable feature, it refers to indexing special or unusual events such as holidays [29].

Finally, it is important to note that forecasting using RNN models is suitable for only one-step ahead. In order to use a multi-step-ahead forecasting, it is important to use an iterative or recursive manner. This latter step can be explained by looping over the RNN model for each new point forecast, and including each new prediction at the beginning of each new loop iteration [29].

3. Results & Discussion

3.1. Data Imputation using the ARIMA Model

The first step of this section is to provide a description of the missing values within the data set. The used BOD5 variable provided by the Chefchaouen WWTP is daily and ranges from the period of the 9th Mai, 2017 to the 28th of February 2018. This accounts for 296 observations where 178 are missing, which represents a percentage of 60.1% of the total observations.

Figure 6 provides an overview of the occurrence of gap sizes, which is the number of NAs in a row, in the data set. This is given such as 3 NAs in a row occurred 6 times, 4 NAs in a row occurred 27 times, 5 NAs occurred 4 times, 6 NAs in a row occurred 1 time, 9 NAs in a row occurred 1 time, and 11 NAs in a row occurred 1 time. The high occurrences of 6 NAs in a row are explained by the fact that the laboratory within this WWTP measures the BOD5 variable only three times a week. For the longer periods of the occurrences of NAs they might be explained by holidays or other special events.

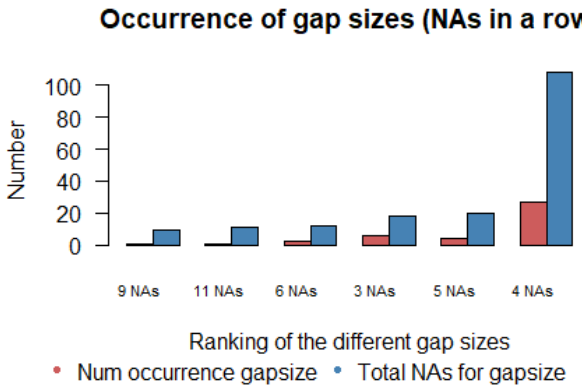

Figure 6. Histogram of the distribution of the missing values gap size

The following line graph (Figure 7) represents the time series BOD5 variable after applying the ARIMA data imputation model.

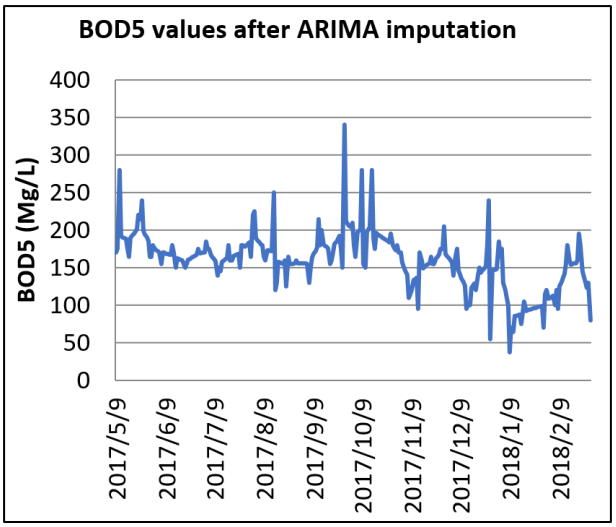

Figure 7. Complete BOD5 values per day using ARIMA data imputation

3.2. BOD5 Variable Forecast using RNN Model

The following section represents the findings of applying the RNN model described in the methods section, and will only introduce the resulting forecasts. Using the BOD5 values of the Chefchaouen WWTPs, two RNN models were trained. Concerning the first trained model, it has used an RNN model with 3 nodes, a learning rate of 0.1, and a number of numepochs of 1000. But concerning the second trained model, it has also used 3 nodes with an additional two hidden layers. Concerning the remaining parameters, they remained the same.

Empirical findings indicate that the first model resulted in an R-squared with a value of 0.58161 and an error term of 10.59 while the second model resulted in an R-squared of 0.59321 and an error term of 9.63. For this the second model will be used to generate predictions.

While the training and testing period covered the period between the 9th of May, 2017 and 4th of January, 2018, the BOD5 forecasts using the iterative model described under the method section are applied to the period between the 5th of January, 2018, and the 28th of February, 2018. This is represented in the following line graph (Figure 8).

The mean absolute error (MAE) was calculated between the BOD5 actual values and the forecasted values using the RNN model 2. This has resulted in a value of 22.09.

3.3. Impact of Predicting BOD5 Variable on Energy Consumption and Oxygen Injection

In order to assess the impact of predicting the BOD5 variable on energy consumption and oxygen injection, the first step is to convert the forecasted BOD5 values under the first model and the second model from Mg/L to Kg/day using the WWTP's water influent and effluent dataset. The second step is to convert each daily observation of Kg of BOD5 per day to the suggested energy consumption and oxygen injection based on their corresponding ratios retrieved from the literature review. These ratios are given such as:

- Since the Chefchaouen WWTP is of type C, meaning that it has a basic treatment of removing carbonaceous pollution and a portion of suspended solids, the energy requirement is assumed to be 1.2 kWh/kg BOD5 (Figure 9) [22].

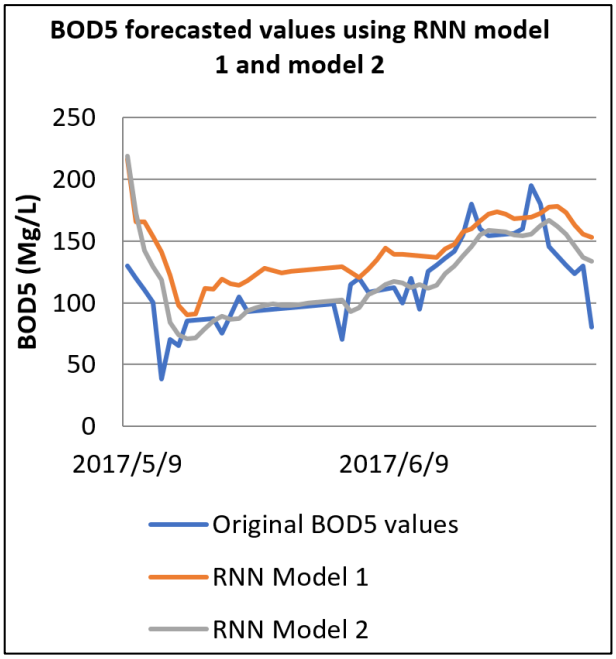

Figure 8. BOD5 forecasted values using RNN model 1 and model 2

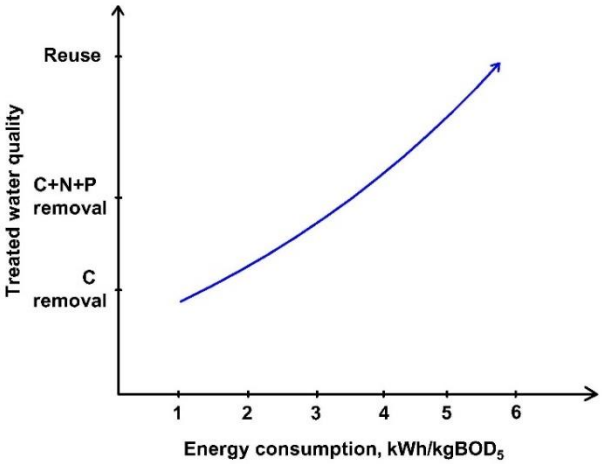

Figure 9. Energy consumption in WWTPs for different levels of treatment for WWTPs in France

- Concerning the average oxygen needed to treat 1 kg of BOD5, it is assumed to be 1.5 kg [9]
- In order to compare the energy consumption, the oxygen consumption and their corresponding suggestions based on the RNN model 1 and model 2, the selected data only accounts for the periods with

actual values of the BOD5 and neglect the periods covered by the ARIMA data imputation method.

Figure 10 shows that the energy consumption in the Chefchaouen WWTP ranges between 707 and 1077 kWh per day with a mean of 891. But concerning the suggested energy consumption by the RNN model 1 and the RNN model 2, it ranges between the values of 348 and 1040, and the values of 262 and 842, with the means of 675 and 563, respectively.

The t-statistic test was conducted in order to compare the mean of the actual energy consumption and the suggested RNN model 1 energy consumption, and in order to compare the mean of the actual energy consumption and the suggested RNN model 2 energy consumption in the period of comparison. Results indicate that the first t-statistic test resulted in a value of 4.53 which corresponds to a p-value of 5.48E-05, and the second t-statistic test resulted in a value of 7.16 which corresponds to a p-value of 4.13E-08. Both of these p-values are under the significance level αequals 5%, which concludes that there is enough evidence that the means are different.

In the periods of comparison, the suggested energy consumption by the first RNN model leads to a decrease of energy by 24.27%, while the suggested energy consumption by the second RNN model leads to a decrease of energy by 36.87%.

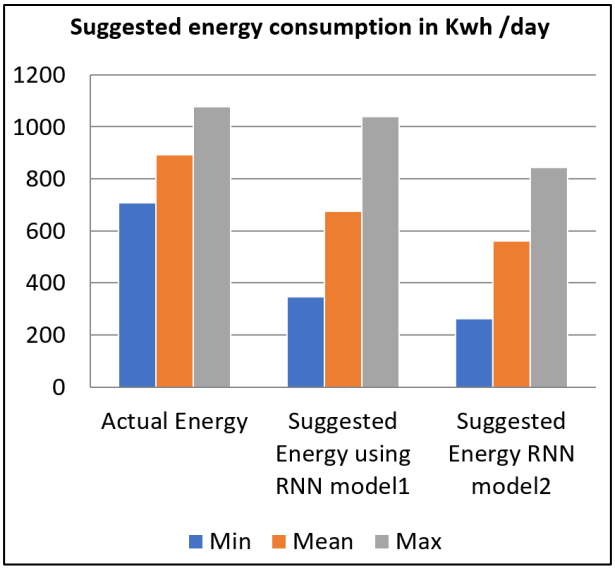

Figure 10. Comparison between actual energy consumption and suggested energy consumption in kWh/day

Concerning Figure 11, it is shown that the actual oxygen injected in kg per day ranges between the value of 5033 and 9347 kg per day, with a mean of 6722. But for the suggested injected oxygen from RNN model 1, it ranges between the values of 434 and 1300, with a mean of 843 kg per day while the suggested injected oxygen from model 2 ranges between the values of 328 and 1053 kg per day, with a mean of 703. This indicates that the Chefchaouen WWTP can maintain a high BOD5 effluent quality while reducing the use of the injected oxygen by 87% and 90% according to the first and second RNN models, respectively.

In order to indicate the difference between the actual oxygen injected mean and the ones suggested by the RNN model 1 and 2, the t-statistic test was used. Concerning the difference between the actual and the RNN model 1 suggested oxygen injection, the t-statistics have resulted in a value of 21.55 which corresponds to a p-value of 4.20E-16. But for the difference between the actual and the RNN model 2 suggested oxygen injection, the t-statistics have resulted in a value of 22.11 which corresponds to a p-value of 7.91E-16.

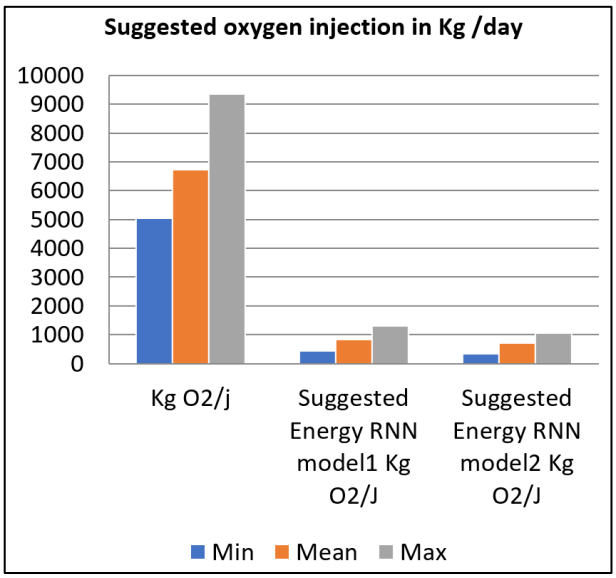

Figure 11. Comparison between actual oxygen injected and suggested oxygen to be injected in kg/day

4. Conclusions

The BOD5 data provided by the management team of the plant had many missing values, as this plant measures this wastewater influent quality parameter only three times per week. For this, the first part of this contribution applied the ARIMA data imputation model in order to complete the dataset that will enable conducting further analyses. Concerning the second step, it relates to the application of the RNN model throughout a trained model (model 1), which performance was further enhanced (model 2).

Empirical findings indicate that the forecasted BOD5 values were close to the actual provided values. The mean absolute error had a value of 22.

In addition to applying the data imputation model and the forecasting model, this contribution assesses the impact of using these predictions. In order to make this assessment, the RNN model has used the period between the 9th of May, 2017 and the 4th of January 2018 to train and test the model, and produced predictions for the period between the

5th of January and 28th of February of year 2018. Using the water inflow, water outflow, and BOD5 effluent data, BOD5 predictions were converted to account for the number of kilograms of BOD5 treated per day in order to suggest the total energy consumption per day as well as the total oxygen to be injected based on ratios retrieved from the existing literature.

Analyses indicate that by applying the ARIMA data imputation model and the RNN model on the Chefchaouen plan, energy consumption could have decreased in the period of assessment by a percentage of 37% while the oxygen injected could have decreased by a percentage of 90%.

REFERENCES

[1] Abba, S. I., Nourani, V., & Elkiran, G.,"Multi-parametric modeling of water treatment plant using AI-based non-linear ensemble," Journal of Water Supply: Research and Technology-Aqua., 2019. DOI: 10.2166/aqua.2019.078

[2] Abunama, T., & Othman, F., "Time series analysis and forecasting of wastewater inflow into Bandar Tun Razak sewage treatment plant in Selangor, Malaysia," IOP Publishing, In IOP conference series: materials science and engineering, vol. 210, no. 1, pp. 012028, 2017. DOI: 10.1088/1757-899X/210/1/012028

[3] Adhikari, R., & Agrawal, R. K.,"An introductory study on time series modeling and forecasting," arXiv preprint arXiv:1302.6613, 2013.

[4] Adib Mashuri, Nur Hamiza Adenan, Nor Suriya Abd Karim, Mohd Shahriman Adenan, Nurulhuda Che Abd Rani, "Performance of Water Level Forecasting Based on Chaos Approach Using Data Splitting," Environment and Ecology Research, Vol. 10, No. 2, pp. 218 - 224, 2022. DOI: 10.13189/eer.2022.100211.

[5] Al-Asheh, S., Mjalli, F. S., & Alfadala, H. E., "Forecasting Influent-Effluent Wastewater Treatment Plant Using Time Series Analysis and Artificial Neural Network Techniques," Chemical Product and Process Modeling, vol. 2, no. 3, 2007. DOI: 10.2202/1934-2659.1063.

[6] Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q., & Zhou, P., "Influent forecasting for wastewater treatment plants in North America," Sustainability, vol. 11, no. 6, pp. 1764, 2019. DOI: 10.3390/su11061764

[7] Che, J., & Wang, J., "Short-term load forecasting using a kernel-based support vector regression combination model," Applied energy, vol. 132, pp. 602-609, 2014. DOI: 10.1016/j.apenergy.2014.07.064

[8] Dyson, B., & Chang, N. B., "Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling," Waste management, vol. 25, no. 7, pp. 669-679, 2005. DOI: 10.1016/j.wasman.2004.10.005

[9] EPA., "Principles of design and operations of wastewater treatment pond systems for plant operators, engineers, and managers," Cincinnatti, OH: United States Environmental Protection Agency, Office of Research and Development., 2011.

[10] Erdem, E., & Shi, J., "ARMA based approaches for forecasting the tuple of wind speed and direction," Applied Energy, vol. 88, no. 4, pp.1405-1414, 2011. DOI: 10.1016/j.apenergy.2010.10.031

[11] Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.)., "Predictive modeling applications in actuarial science," Cambridge University Press., vol. 1, 2014.

[12] Harrison, P. J., & Stevens, C. F., "Bayesian forecasting," Journal of the Royal Statistical Society: Series B (Methodological), vol. 38, no. 3, pp. 205-228, 1976. http://www.jstor.org/stable/2984970

[13] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," Neural computation, vol. 9, no. 8, pp.1735–1780, 1997.

[14] Johnston, F. R., Boyland, J. E., Meadows, M., & Shale, E., "Some properties of a simple moving average when applied to forecasting a time series," Journal of the Operational Research Society, vol. 50, no. 12, pp.1267-1271, 1999. http://www.jstor.org/stable/3010636

[15] Khwaja, A. S., Anpalagan, A., Naeem, M., & Venkatesh, B., "Joint bagged-boosted artificial neural networks: Using ensemble machine learning to improve short-term electricity load forecasting," Electric Power Systems Research, 179, 106080, 2020. DOI : 10.1016/j.epsr.2019.106080

[16] Kim, J. R., Ko, J. H., Im, J. H., Lee, S. H., Kim, S. H., Kim, C. W., & Park, T. J., "Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model," Water science and technology, vol. 53, no. 4-5, pp.185-192, 2006. DOI : 10.2166/wst.2006.123

[17] Kim, M., Kim, Y., Kim, H., Piao, W., & Kim, C., "Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant," Frontiers of Environmental Science & Engineering, vol.10, no.2, pp. 299–310, 2015. DOI: 10.1007/s11783-015-0825-7.

[18] Kim, T. K., "T test as a parametric statistic," Korean Journal of Anesthesiology, vol. 68, no. 6, pp. 540, 2015. DOI: 10.4097/kjae.2015.68.6.540.

[19] Kim, H. S., Kim, Y. J., Cheon, S. P., Baek, G. D., Kim, S. S., & Kim, C. W., "Evaluation of model-based control strategy based on generated setpoint schedules for NH4–N removal in a pilot-scale A2/O process," Chemical Engineering Journal, vol. 203, pp. 387–397, 2012. DOI: 10.1016/j.cej.2012.07.067

[20] Kriger, C., & Tzoneva, R., "Neural networks for prediction of wastewater treatment plant influent disturbances," AFRICON, 2007. DOI :10.1109/afrcon.2007.4401646.

[21] Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D., "Judgmental forecasting: A review of progress over the last 25 years," International Journal of forecasting, vol. 22, no. 3, pp. 493-518, 2006. DOI: 10.1016/j.ijforecast.2006.03.007

[22] Lazarova, V., Choo, K.-H., & Cornel, P., "Water-energy interactions in water reuse," London: IWA., 2012.

[23] Lewis, N. D. C., "Neural Networks for Time Series Forecasting with R: an Intuitive Step by Step Blueprint for Beginners," AusCov., 2017.

[24] Li, G., Song, H., & Witt, S. F., "Recent developments in econometric modeling and forecasting," Journal of Travel Research, vol. 44, no.1, pp. 82-99, 2005.

[25] Lotfi, K., Bonakdari, H., Ebtehaj, I., Mjalli, F. S., Zeynoddin, M., Delatolla, R., & Gharabaghi, B., "Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology," Journal of Environmental Management, vol. 240, pp. 463–474, 2019. DOI: 10.1016/j.jenvman.2019.03.137

[26] Melo, J. D., Carreno, E. M., & Padilha-Feltrin, A., "Multi-agent simulation of urban social dynamics for spatial load forecasting," IEEE Transactions on Power Systems, vol. 27, no.4, pp. 1870-1878., 2012. DOI: 10.1109/TPWRS.2012.2190109

[27] Nadiri, A. A., Shokri, S., Tsai, F. T. C., & Moghaddam, A. A., "Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model," Journal of Cleaner Production, vol. 180, pp. 539-549., 2018. DOI : 10.1016/j.jclepro.2018.01.139

[28] Ostertagova, E., & Ostertag, O., "Forecasting using simple exponential smoothing method," Acta Electrotechnica et Informatica, vol. 12, no. 3, pp. 62, 2012. DOI : 10.2478/v10198-012-0034-2

[29] Petneházi, G., "Recurrent neural networks for time series forecasting," arXiv preprint arXiv:1901.00069, 2019.

[30] Reagan, K. M., "An evaluation of ARIMA(Box-Jenkins) models for forecasting wastewater treatment process variables," Master's thesis, UCLA, 1984.

[31] Schmidberger, T., Posch, C., Sasse, A., Gülch, C., & Huber, R., "Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling," Biotechnology progress, vol. 31, no. 4, pp. 1119-1127, 2015. DOI: 10.1002/btpr.2105

[32] Shahidah Othman, Rosnalini Mansor, Fakhrurazi Ahmad, "Weighted Subsethood Fuzzy Time Series towards Energy-Water Efficiency for Water Treatment Plant," Environment and Ecology Research, Vol. 10, No. 2, pp. 182 - 192, 2022. DOI: 10.13189/eer.2022.100207.

[33] Shakti, S. P., Hassan, M. K., Zhenning, Y., Caytiles, R. D., & SN, I. N. C. , "Annual Automobile Sales Prediction Using ARIMA Model," Int. J. Hybrid Inf. Technol, vol. 10, pp. 13-22, 2017. DOI: 10.14257/ijhit.2017.10.6.02

[34] Wang, G. C., & Jain, C. L., "Regression analysis: modeling & forecasting," Institute of Business Forec., 2003.

[35] Yu, T., Yang, S., Bai, Y., Gao, X., & Li, C., "Inlet water quality forecasting of wastewater treatment based on kernel principal component analysis and an extreme learning machine," Water, vol. 10, no. 7, pp. 873, 2018. DOI: 10.3390/w10070873

[36] Zhou, Y., Huang, G., Zhu, H., Li, Z., & Chen, J., "A factorial dual-objective rural environmental management model," Journal of Cleaner Production, vol. 124, pp. 204-216, 2016. DOI : 10.1016/j.jclepro.2016.02.081

[37] Zhou, Y., Yang, B., Han, J., & Huang, Y., "Robust linear programming and its application to water and environmental decision-making under uncertainty," Sustainability, vol. 11, no. 1, pp. 33, 2019. DOI : 10.3390/su11010033