Integrating Fuzzy System and Meta-Heuristic Algorithms to Predict Influent Parameters for a Sewage Treatment

Mozafar Ansari¹, Faridah Othman*¹ and El-Shafie¹

¹ Department of Civil Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, MALAYSIA
*Corresponding author: faridahothman@um.edu.my

Abstract. Sewage treatment plants (STPs) are built to reduce the concentration of sewage parameters to a safe level that reduced their impact on the environment. To have an optimal operation of STPs, it is essential to estimate influent parameters precisely. In this research, four influent physicochemical characteristics, i.e. biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH₃-N), and suspended solids (SS), of a sewage treatment plant (STP) were analysed and predicted by integrated genetic algorithm Sugeno fuzzy inference system (GA-FIS) and particle swarm optimisation FIS (PSO-FIS). The GA-FIS and PSO-FIS were applied on 10 time-series scenarios, and the results of each scenario were compared to find the best algorithm as well as the scenario for each parameter. Based on the results, both GA-FIS and PSO-FIS algorithms provided very good results, and the differences of error in predicting influent parameters is very less. However, to select the best algorithm for predicting the missing values of each parameter, GA-FIS predicted BOD and SS more accurately than PSO-FIS algorithm, COD and ammoniacal nitrogen had more accurate results when they were predicted by PSO-FIS algorithm.

1. Introduction

Development of metropolitan has a severe threat to the environment and put many inhabitant’s life in danger. One of these threats is emerging effluent of industrial and municipal to waterbody which increases the contamination of water. To control the pollution of water, sewage treatment plants (STPs) have been developed in order to reduce the concentration of sewage before discharging to the environment. The treatment process contains a considerable number of complex physical, chemical and biological processes. Minimising the cost of treating sewage is very important for scientist, operators, and planner and different parameters such as influent characteristics, have an influence on the treatment process. Therefore, it is necessary to predict influent parameters accurately.

Researchers are developing more robust algorithms to have a better prediction result. In hydrological modelling, before artificial intelligence (AI) algorithms become very popular, statistical methods such as Autoregressive Integrated Moving Average (ARIMA), were used to predict their data [1]. However, due to prediction accuracy of AI algorithms, many hydrological researchers have used black-box data-mining methods such as an artificial neural network (ANN), support vector machine (SVM), and adaptive neural fuzzy inference system (ANFIS), in order to get more precise results [2-6].

These black-box data-mining algorithms have their benefits and drawbacks. For example, SVM has difficulty in dealing with large data sets and needed to be integrated with another algorithm to improve its prediction accuracy [7]. Therefore, AI algorithms can be integrated with other algorithms such as genetic algorithms (GA), particle swarm optimisation (PSO), wavelet, etc., to deliver better results [8-
10]. Najah Ahmed et al. [11] applied different data-mining techniques to predict several water quality parameters. They found that the overall prediction accuracy of a wavelet de-noising technique with the Neuro-Fuzzy Inference System predicted (WDT-ANFIS) was better than other algorithms that were used in their study.

Several algorithms have been used to evaluate the accuracy of influent parameters [12-13]. Ansari et al. [14] compared different time-series algorithms, viz. ARIMA, multilayer perceptron (MLP), and SVM, to predict influent flow-rate. They found that both MLP and SVM models had better prediction results than ARIMA method.

In this research, a Sugeno fuzzy inference system (FIS) algorithm integrated with two optimisation algorithms, i.e. GA and PSO, were used to predict four well-known influent parameters, biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH₃-N), and suspended solids (SS), within different time-series scenarios. The results of all scenarios were compared to find the best algorithm for each parameter.

2. Methods and Materials

2.1. Study area

In this research, four physicochemical characteristics of a Sewage treatment plant (STP) were analysed. The STP is located at Jalan Damansara, Kuala Lumpur. It was constructed to treat sewage by using oxidation ditches process. The daily capacity is 100,000 PE, and the daily flow is 25,000 [15].

Three years of weekly influent parameters, viz. BOD, COD, NH₃-N, and SS, from 2011 to 2013 were used in this research. Figure 1 shows the available records.

![Figure 1](image-url)

**Figure 1.** STP influent data, (a) BOD, (b) COD, (c) NH₃-N, and (d) SS, from 2011 to 2013.

2.2. Processing data

In this research, two integrated fuzzy systems were developed to model the data. However, before applying developed algorithms on the influent data, it is necessary to run an outlier detection test to
highlight unrealistic records. A developed function in MATLAB 2017b found outliers records using Grubbs Test [16] and replaced them with a proper value.

2.2.1. Input of model

Ten time-series scenarios were constructed to compare different model input sets and find optimal results in forecasting influent parameters. Table 1 summarises input scenarios for each quality parameter. Due to missing values, only records that have minimum uninterrupted records regarding its scenario were selected in the time-series model. In each algorithm, 80% of input data were selected randomly for training algorithm and results were evaluated based on 100% of data.

| Scenario | Input                        | Output |
|----------|------------------------------|--------|
| 1        | Q(t), Q(t-1), P(t-1)         | P(t)   |
| 2        | Q(t), Q(t-1), Q(t-2), P(t-1), P(t-2) | P(t)   |
| 3        | Q(t), Q(t-1), ..., Q(t-3), P(t-1), ..., P(t-3) | P(t)   |
| 4        | Q(t), Q(t-1), ..., Q(t-4), P(t-1), ..., P(t-4) | P(t)   |
| 5        | Q(t), Q(t-1), ..., Q(t-5), P(t-1), ..., P(t-5) | P(t)   |
| 6        | Q(t), Q(t-1), ..., Q(t-6), P(t-1), ..., P(t-6) | P(t)   |
| 7        | Q(t), Q(t-1), ..., Q(t-7), P(t-1), ..., P(t-7) | P(t)   |
| 8        | Q(t), Q(t-1), ..., Q(t-8), P(t-1), ..., P(t-8) | P(t)   |
| 9        | Q(t), Q(t-1), ..., Q(t-9), P(t-1), ..., P(t-9) | P(t)   |
| 10       | Q(t), Q(t-1), ..., Q(t-10), P(t-1), ..., P(t-10) | P(t)   |

2.2.2. Fuzzy inference system (FIS)

In this research, the Sugeno fuzzy system was used to model the data. Between subtractive clustering (SC) and grid partitioning (GP), as the conventional methods in generating Sugeno fuzzy inference system (FIS), the SC method does not require computational effort as much as GP method. Therefore, subtractive clustering technique was utilised in forecasting influent parameters. The SC method has been developed by Chiu [17], and in this method, each point is considered as a potential cluster centre, and then based on the density of the adjacent data point, the potentiality of each data point will be measured [18]. The fuzzy rules in the Takagi and Sugeno fuzzy inference system [19] were represented in equation 1.

\[
\text{if } \text{X is A}_i \text{ and } \text{Y is B}_i \text{ then } Z_i = p_i X + q_i Y + r_i \quad \forall i = 1, ..., n
\]  

(1)

Where \(p_i\), \(q_i\), \(r_i\) are the Sugeno fuzzy parameter set.

In this research, 37 different cluster influence ranges (CIR) from 0.05 to 0.95 were assigned to develop the Sugeno FIS. Moreover, the membership function (MF) ranges of the FIS were adjusted between -10 to 10 by using genetic algorithms (GA) and particle swarm optimisation (PSO) algorithms.

2.2.3. Integrated Genetic Algorithm FIS

Genetic algorithm is a stochastic and population-based optimisation algorithm. The idea of the GA technique has been inspired by natural selection, mutation, crossover, and inheritance as natural phenomena [20-21]. The GA has been developed by Holland [22], and it seeks optimum results by evaluating and choosing parent population generated population using the roulette-wheel method and by applying mutation, crossover operators on the parent to produce the next generation in each iteration. The fundamentals of the genetic algorithm are comprehensively explained by Goldberg [23].
2.2.4. Particle swarm optimisation FIS

One of the most famous swarm intelligence algorithms is Particle Swarm Optimization. Particle swarm optimisation (PSO) is a population-based and stochastic optimisation technique that inspired the intelligence behaviour of swarms such as birds or fish [18]. The PSO model developed by Kennedy and Eberhart [24]. Unlike evolutionary algorithms, such as genetic algorithm, PSO does not use any evolution operator, like mutation or crossover, and selection method. The population survive from the beginning until the end of the trial [20, 25].

Particle swarm optimisation (PSO) finds the optimum result based on the best particle’s position. PSO starts its algorithm by generating a random population, which calls particles, in a space of solutions. The population locations were evaluated by cost functions such as MSE, RMSE, MAE, etc. Then, the best local and best global positions were defined, and each particle moves in the limited space of solution with a specified velocity, to find the optimal solution.

PSO has been applied to train the MF parameters of the initial fuzzy model to reduce the model error. PSO algorithm can operate well with a small population. In each iteration, the best local and global results have been defined by RMSE.

2.3. Evaluation criteria

In this research, two indices were used to find the accuracy of the models’ result. For all parameters, root mean square error (RMSE) value was calculated to find the average error of algorithms’ results and select the best cluster influence range (CIR) value in each scenario. Also, the RMSE is used to compare the results of each scenario and selecting the best scenario for each parameter. The RMSE is calculated by equation 2.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_{prd,i} - P_{obs,i})^2}{N}} \tag{2}
\]

Where N is the size of data, Pobs, i is the ith observed value of an influent quality parameter, \(P_{prd,i}\), is predicted value and \(\bar{P}_{obs}\), \(\bar{P}_{prd}\) are average of observed and predicted values.

In addition, the coefficient of determination (R\(^2\)) is calculated to compare and validate the results. Coefficient of determination ranges is from 0 to 1, the higher value of R\(^2\) shows better agreement between observed and simulated results. In the hydrological studies, values less than 0.5 is considered as poor and unacceptable results, between 0.5 to 0.6 is satisfactory, any value ranges 0.6 to 0.7 is good, and values higher than 0.7 indicates very good results [14, 26].

\[
R^2 = \left(\frac{\sum_{i=1}^{N}(P_{obs,i} - \bar{P}_{obs})(P_{prd,i} - \bar{P}_{prd})}{\sum_{i=1}^{N}(P_{obs,i} - \bar{P}_{obs})^2 \sum_{i=1}^{N}(P_{prd,i} - \bar{P}_{prd})^2}\right)^2 \tag{3}
\]

3. Discussion and Results

All unreliable records were highlighted by using Grubbs Test and replaced by a proper value. When reliable values of previous and subsequent an outlier were available, then the outlier was replaced with the average of them. Otherwise, it replaced with the total average of the series. Based on the results, four records were detected as an outlier for BOD and COD parameters, and two records for \(\text{NH}_3\)-N and SS, totally 12 records.

3.1. Model Results

The RMSE and R\(^2\) of each scenario were considered to evaluate and select the best scenario of each parameter. The results of each algorithm were presented separately.
3.1.1 Integrated GA-FIS

A real-code GA was used to minimise the fuzzy model errors. Because each scenario was repeated with 37 different CIRs, it was not possible to represent detailed results. Therefore, the results of GA-FIS with best CIR presented in Table 2.

| Scenario | BOD       | COD       | NH$_3$-N | SS        |
|----------|-----------|-----------|----------|-----------|
|          | RMSE      | $R^2$     | RMSE     | $R^2$     | RMSE     | $R^2$     | RMSE     | $R^2$     |
| 1        | 22.2360   | 0.7594    | 70.2709  | 0.5861    | 2.1414   | 0.7846    | 17.9125  | 0.8692    |
| 2        | 27.1813   | 0.6134    | 59.2179  | 0.6460    | 2.2010   | 0.7594    | 28.5914  | 0.6945    |
| 3        | 18.8054   | 0.8081    | 34.6348  | 0.8590    | 2.4902   | 0.6698    | 31.9513  | 0.5941    |
| 4        | 8.5135    | 0.9596    | 59.9172  | 0.6533    | 2.9461   | 0.5662    | 34.1473  | 0.5794    |
| 5        | 18.7595   | 0.8094    | 41.1820  | 0.8165    | 1.9439   | 0.8147    | 12.5702  | 0.9443    |
| 6        | 19.4703   | 0.7950    | 49.5761  | 0.7625    | 2.8723   | 0.6109    | 21.0587  | 0.8540    |
| 7        | 22.4709   | 0.7296    | 40.0366  | 0.8544    | 1.3913   | 0.9059    | 26.8506  | 0.7733    |
| 8        | 17.6816   | 0.8401    | 64.4228  | 0.6103    | 1.3823   | 0.9006    | 32.0512  | 0.6243    |
| 9        | 11.4543   | 0.9335    | 47.6629  | 0.8992    | 2.5566   | 0.6849    | 16.5325  | 0.9078    |
| 10       | 26.0868   | 0.6717    | 47.6629  | 0.8992    | 1.6555   | 0.8982    | 16.8229  | 0.9040    |

Based on the GA-FIS algorithm results for BOD, scenarios 2 and 10 were in a good range, and other scenarios had very good results by considering the value of $R^2$. Among these scenarios, scenario 4 had a minimum error and the highest $R^2$. The results of COD showed that scenario 1 had satisfactory results, scenarios 2, 4, and 8 had good results, and other scenarios had very good agreement between observed values and predicted values. Moreover, scenario 3 had the minimum RMSE and the highest coefficient of determination. Comparing the results of NH$_3$-N scenarios, scenarios 1, 2, 5, 7, 8, and 10 had very good results, and the rest had either good or satisfactory results. Although the scenario 8 had a lesser value of $R^2$ than scenario 7, scenario 8 was chosen as the best results because it had less error than scenario 7. The results of SS scenarios showed that scenario 5 had the highest $R^2$ and the least RMSE. Therefore, scenario 4 in BOD, scenario 3 in COD, scenario 8 in NH$_3$-N, and scenario 5 for SS were selected as the best GA-FIS scenarios and compared with the PSO-FIS results.

3.1.2. Integrated PSO-FIS

Membership functions ranges of initial FIS model were optimised by the PSO algorithm. Same as GA-FIS model, the results of PSO-FIS with optimal CIR were shown in Table 3.

The PSO-FIS results in modelling BOD concentration showed that scenario 1 was in a satisfactory range, scenario 3 was in a good range, and the rest scenarios were in a very good range. Among these scenarios, scenario 5 had a minimum error, and scenario 10 had the highest $R^2$. As a result, the difference between $R^2$ is not considerable, and scenario 5 was selected for the best BOD results among PSO-FIS scenarios. The results of COD showed that only scenario 9 had good results, and others had very good results. Besides, scenario 6 had minimum RMSE, and maximum $R^2$ compared the other scenarios. Scenarios 1, 4, and 8 had good prediction results in predicting NH$_3$-N and others had very good results. The results showed that scenario 7 provided a more accurate prediction result than other scenarios. The results of SS scenarios showed that except scenario 3, other scenarios had $R^2$ greater than 0.7, which considered as very good prediction results. Among them, scenario 7 the least RMSE value and the highest $R^2$ value.
### Table 3. PSO-FIS results of modelling the STP influent parameters

| Scenario | BOD RMSE | R² | COD RMSE | R² | NH₃-N RMSE | R² | SS RMSE | R² |
|----------|----------|----|----------|----|------------|----|---------|----|
| 1        | 30.8676  | 0.5633 | 53.2764 | 0.7127 | 2.9423 | 0.6102 | 25.0323 | 0.7576 |
| 2        | 21.8980  | 0.7595 | 39.9300 | 0.8195 | 2.4103 | 0.7122 | 29.1360 | 0.7004 |
| 3        | 23.4404  | 0.6970 | 42.9227 | 0.7841 | 2.1619 | 0.7606 | 27.6091 | 0.6868 |
| 4        | 18.4463  | 0.8115 | 36.1465 | 0.8611 | 2.6572 | 0.6237 | 24.7271 | 0.7887 |
| 5        | 12.2357  | 0.9156 | 39.5674 | 0.8295 | 2.3520 | 0.7016 | 17.7681 | 0.8812 |
| 6        | 20.5225  | 0.7691 | 33.9903 | 0.8896 | 2.2760 | 0.7337 | 21.3469 | 0.8525 |
| 7        | 15.8059  | 0.8668 | 49.9096 | 0.7668 | 0.7115 | 0.9747 | 17.0737 | 0.9020 |
| 8        | 15.8754  | 0.8673 | 47.6974 | 0.8168 | 2.8945 | 0.6035 | 27.0898 | 0.7447 |
| 9        | 16.8316  | 0.8578 | 66.5941 | 0.6164 | 1.8730 | 0.8415 | 18.1447 | 0.8939 |
| 10       | 12.7857  | 0.9196 | 51.5951 | 0.7870 | 1.6887 | 0.8965 | 28.8539 | 0.7199 |

Therefore, scenarios 5, 6, 7, and 7 were nominated as the best scenarios to model BOD, COD, NH₃-N, and SS, respectively, by using PSO-FIS algorithm. Their results were compared by GA-FIS results.

3.2. Predicting influent parameters

Both GA-FIS and PSO-FIS algorithms provided accurate results for each parameter. However, the best scenarios of these algorithms should be compared to select the most precise algorithm to model the STP influent parameters. Therefore, the best results of each algorithm were represented in Table 4 to have easier comparisons.

#### Table 4. The selected scenarios

| Scenario | GA-FIS | PSO-FIS |
|----------|--------|---------|
|         | RMSE   | R²      | RMSE   | R²      |
| BOD     | 4      | 8.513   | 0.960  | 5       | 12.236  | 0.916  |
| COD     | 3      | 34.635  | 0.859  | 6       | 33.990  | 0.890  |
| NH₃-N   | 8      | 1.382   | 0.901  | 7       | 0.711   | 0.975  |
| SS      | 5      | 12.570  | 0.944  | 7       | 17.074  | 0.902  |

Based on Table 4, an integrated genetic algorithm with fuzzy inference system provided more accurate results in predicting the concentration of BOD and SS. PSO, obviously, provided better modelling results for COD and ammoniacal nitrogen concentrations. Therefore, influent parameters of the STP were predicted based on the most fitted algorithms. The prediction results are presented in Figure 2.
4. Conclusions

Having an accurate modelling result of influent parameters, especially in optimising the operation, or planning to develop an STP, is very desirable for scientists, planners, and operators. Influent parameters of an STP are fundamental to operate or establish an STP. Based on the availability of data and expected accuracy, different data-mining techniques can be used to model influent data. Robust prediction algorithms are more complicated and need more computational efforts, and usually, they provide more accurate results. In some cases, the number of available data is not enough to get proper results by using simple AI techniques such as MLP. Therefore, in this research, four physicochemical influent parameters, i.e. BOD, COD, NH3-N, and SS, were modelled using two integrated fuzzy system algorithms and the results of predictions were compared.
The RMSE and $R^2$ results of the study area in this research showed that both GA-FIS and PSO-FIS algorithms could predict the influent parameters very well. GA-FIS had the highest accuracy in predicting BOD and SS concentration, and PSO-FIS had the best prediction results for COD and NH$_3$-N concentration.

5. References

[1] Narayanan P, Basistha A, Sarkar S and Kamna S 2013 Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India Comptes Rendus Geoscience 345(1) 22-7
[2] Shahverdi K, Khodaei K and Shokohi A 2015 Prediction of Monthly Inflow to Dam Reservoir using ANN and ANFIS: A Case Study of Latyan Dam International Bulletin of Water Resources and Development (IBWRD) 3(3) X-XVIII
[3] Yaseen Z M 2015 RBFNN versus FFNN for daily river flow forecasting at Johor River, Malaysia Neural Comput & Appl. 27(6) 1533-1542
[4] Raghavendra. N S, Deka PC 2014 Support vector machine applications in the field of hydrology: A review Applied Soft Computing 19 372-86
[5] Noori R, Karbassi A, Ashrafi K, Ardestani M, Mehrdadi N and Nabi Bidhendi G-R 2012 Active and online prediction of BOD5 in river systems using reduced-order support vector machine Environmental Earth Sciences 67 (1) 141-149
[6] Abobakr Yahya AS, Ahmed AN, Binti Othman F, Ibrahim RK, Afan HA and El-Shafie A, et al. 2019 Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Waters Journal of Hydrology 578 1245-1254
[7] Collobert R and Bengio S 2001 SVMTorch: Support vector machines for large-scale regression problems Journal of machine learning research 1(Feb) 143-60
[8] Ashu J and Sanaga S 2004 Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques Water Resources Research 40(4)
[9] Wei S, Song J and Khan NI 2012 Simulating and predicting river discharge time series using a wavelet-neural network hybrid modelling approach Hydrological Processes 26 (2) 281-96
[10] Barzegar R, Adamowski J and Moghaddam AA 2016 Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran Stochastic Environmental Research and Risk Assessment 30 (7) 1797-819
[11] Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H, Khaleel Ibrahim R, Ming Fai C and Shabbir Hossain M et al. 2019 Machine learning methods for better water quality prediction Journal of Hydrology 578 124084
[12] Ansari M, Othman F and El-Shafie A 2018 Evaluating the Efficiency of Artificial Intelligence Models to Predict Missing Water Quality Parameter Data for a Wastewater Treatment Plant 4th International Conference on Sustainable Civil Engineering Structures and Construction Materials (SCESCM) (Yogyakarta, Indonesia)
[13] Wei X and Kusiak A 2015 Short-term prediction of influent flow in wastewater treatment plant Stochastic Environmental Research and Risk Assessment 29 (1) 241-249
[14] Ansari M, Othman F, Abunama T and El-Shafie A 2018 Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant, Malaysia Environmental Science and Pollution Research 25 (12) 12139-49.
[15] Awang N A and Shaaban MG 2016 Effect of reactor height/diameter ratio and organic loading rate on formation of aerobic granular sludge in sewage treatment International Biodeterioration & Biodegradation 112 (Supplement C) 1-11
[16] Grubbs F E and Beck G 1972 Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations Technometrics 14 (4) 847-54
[17] Chiu SL 1994 Fuzzy Model Identification Based on Cluster Estimation J Intell Fuzzy Syst 2(3) 267-78
[18] Mirrashid M 2014 Earthquake magnitude prediction by adaptive neuro-fuzzy inference system (ANFIS) based on fuzzy C-means algorithm Natural Hazards 74(3) 1577-93
[19] Takagi T and Sugeno M 1985 Fuzzy identification of systems and its applications to modeling and control IEEE Transactions on Systems, Man, and Cybernetics SMC-15 (1) 116-32
[20] Rani D and Moreira M M 2010 Simulation–Optimization Modeling: A Survey and Potential Application in Reservoir Systems Operation Water Resources Management 24(6) 1107-38
[21] Al-Obaidi M A, Li J P, Kara-Zaïtri C and Mujtaba IM 2017 Optimisation of reverse osmosis based wastewater treatment system for the removal of chlorophenol using genetic algorithms Chemical Engineering Journal 316 91-100
[22] Holland J H 1975 Adaptation in Natural and Artificial Systems (Ann Arbor, MI: The University of Michigan Press)
[23] Goldberg D E 1989 Genetic algorithms in search, optimization, and machine learning Addison-Wesley Pub Co 412
[24] Kennedy J and Eberhart R 1995 Particle Swarm Optimization IEEE international conference on neural networks (Perth, Australia) 1942–8
[25] Kennedy J 2010 Particle Swarm Optimization Encyclopedia of Machine Learning (Boston, MA: Springer US) 760-6
[26] Ayele G T, Teshale E Z, Yu B, Rutherfurd I D and Jeong J 2017 Streamflow and sediment yield prediction for watershed prioritization in the upper Blue Nile river basin, Ethiopia Water (Switzerland) 9(10).

Acknowledgments
The authors would like to acknowledge The Ministry of Education and the University of Malaya for their financial support through grants FRGS (FP016-2014A) and UMRG (FL001-13SUS). We would also like to thank the cooperation of University of Malaya Water Engineering and Spatial-Environment Governance Research Centre (WESERGE) and other relevant companies and parties for their assistant in conducting this research. We highly appreciate and would like to thank the reviewers for their valuable suggestions.