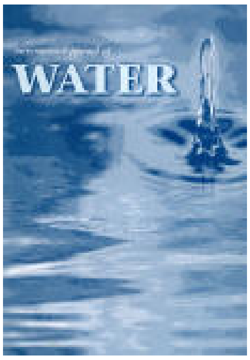
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Water

ISSN online: 1741-5322 - ISSN print: 1465-6620
https://www.inderscience.com/ijw

Predicting aluminium using full-scale data of a conventional water treatment plant on Orontes River by ANN, GEP, and DT

Ruba Dahham Alsaeed, Bassam Alaji, Mazen Ibrahim

DOI: 10.1504/IJW.2023.10058775

Article History:

Received:	12 April 2022
Accepted:	04 April 2023
Published online:	09 October 2023

Predicting aluminium using full-scale data of a conventional water treatment plant on Orontes River by ANN, GEP, and DT

Ruba Dahham Alsaeed*

Faculty of Engineering,
Al-Wataniya Private University,
Hama, Syria
ORCID: 0000-0002-4848-5348
Email: rubaalsaeed89@gmail.com
*Corresponding author

Bassam Alaji

Faculty of Civil Engineering,
Department of Sanitary and Environmental Engineering,
Damascus University,
Damascus, Syria
Email: bassam.eng.77.env@gmail.com

Mazen Ibrahim

Faculty of Civil Engineering,
Department of Engineering Management and Construction,
Damascus University,
Damascus, Syria
Email: mazen.ibrahim@damascusuniversity.edu.sy

Abstract: Aluminium sulphate is one of the most common chemicals used to coagulate water. Some studies indicate that it can increase the risk of Alzheimer's disease. This study focused on the relationship between residual aluminium and many parameters. The actual data of Al-Qusayr purification plant in Homs city was used. Three different models were studied, artificial neural networks (ANN), genetic expression technology (GEP) and Decision Tree (DT), to determine the residual aluminium. The models' results were compared. ANN was the best in modelling data when initial turbidity was between 6.5 and 30 NTU, decision tree was better in the range 25 to 60 NTU. In general the best model was ANN, while the most easily generalised one was GEP. The ANN model was found to be the most suitable model to predict residual aluminium with a coefficient of determination $R^2 = 0.88$ and RMSE = 0.019 mg/L.

Keywords: aluminium residual; artificial neural networks; gene expression; decision tree; turbidity.

Reference to this paper should be made as follows: Alsaeed, R.D., Alaji, B. and Ibrahim, M. (2023) 'Predicting aluminium using full-scale data of a conventional water treatment plant on Orontes River by ANN, GEP, and DT', *Int. J. Water*, Vol. 15, No. 3, pp.190–206.

Biographical notes: Ruba Dahham Alsaeed, Doctor in Environmental Engineering. She taught mechanical engineering, descriptive geometry, and environmental engineering at Albaath University from 2014 to 2021, and taught mechanical engineering at Al-Wataniya Private University from 2020 to 2022. She works as an Assistant Professor at Engineering College in Al-Wataniya Private University, teaching drinking water distribution systems, waste water treatment and drinking water purification.

Bassam Alaji, Member of academic staff at the department of sanitary and environmental engineering – Faculty of civil Engineering – Damascus University, since 1995. He is Professor since July 2017. He is Head of Sanitary and Environmental Engineering Department – Faculty of civil Engineering – Damascus University, since September 2016 until September 2020. Teaching: water treatment, ecology and environmental engineering, environmental protection, solid waste management, water and wastewater net and advanced industrial water treatment.

Mazen Ibrahim, Assistant Professor at the Faculty of Civil Engineering, Department of Engineering Management and Construction, Damascus University.

1 Introduction

Rivers and lakes are among the most important sources of drinking water. These water sources are polluted by various sources of pollution (Alsaeed et al., 2022a). Water purification requires specific attention to meet the standards required (Tahraoui et al., 2021). The main objective of the purification plant is to produce drinking water that is safe for consumption; that does not contain pathogenic or toxic agents, and this must be done at the lowest possible cost, and with the least impact on the environment.

Surface water contains suspended matter with a specific gravity greater than one. Suspended substances tend to settle to the bottom of the waterbed, but fine particles of small dimensions remain in the water in the suspended state (Amin and Sadaf, 2018). Coagulation is the process of neutralisation of colloidal particles by adding a chemical coagulant or conditioning process to enhance their agglomeration and thus produce larger particles that can be removed more easily in subsequent processing operations.

When coagulant is added to raw water, positively charged coagulants react with dissolved particles and colloids, in coagulants that seek to destabilise the particles by neutralising the charge the necessary dose of coagulant will have a turbidity relationship (Krupińska, 2020).

The most commonly used coagulants in drinking water treatment are aluminium coagulants, aluminium sulphate ($Al_2(SO_4)_3.18H_2O$) or what is known commercially as alum. The coagulant reacts with the alkaline present such as carbonate, bicarbonate, hydroxide or phosphate to form insoluble aluminium salts.

Aluminium is an amphoteric compound, which combines with both acids and bases to form, respectively, aluminium salts and aluminates. The chemical presence of aluminium in water is mainly Al $(OH)_3$ which has an amphoteric character and a tendency to form complex ions (Krupińska, 2020).

Since aluminium is added to the water purification process, the aluminium value in treated water is often higher than in raw water. Therefore, the remaining aluminium is related to the processing process and is used to evaluate the performance of the process . (Kim and Yoon, 2000)

The use of aluminium coagulants in drinking water treatment leads to high concentrations of aluminium in drinking water. High concentrations of aluminium may increase the turbidity of the water in the distribution system by precipitating aluminium hydroxide. The $A1^{+3}$ ion forms strong bonds with oxygen. This weakens the bonding of the oxygen and hydrogen atoms in water molecules, and the hydrogen atoms tend to be freed in solution.

This process is known as hydrolysis, and the resulting aluminium hydroxide species are called hydrolysis products.

The chemistry of the reactions and products of aluminium hydrolysis is complex and not fully understood. Hydrolysis products tend to adsorb (and may continue to hydrolyse). The form of precipitated aluminium depends on the conditions of formation, including temperature, and the pH of the solution.

Aluminium ions complement reactions in the human body with metal ions such as zinc, iron, calcium and chromium. Once absorbed, the aluminium reaches the blood and is mainly transferred to transferrin and can cross the blood-brain barrier.

Symptoms of nausea, vomiting and diarrhoea have been reported at high levels of aluminium residue in drinking water, as well as mouth and skin ulcers, rash and joint pain (Tomperi et al., 2013)

The concentration of aluminium in water can vary greatly depending on different physical and chemical substances and mineral factors.

The Environmental Protection Organization has stated the permissible limit for aluminium (0.05–0.2 mg/L) and the World Health Organization is 0.2 mg/L.

The artificial intelligence sector is witnessing a continuous development, making it a safe haven for environmental and natural resource management experts in search of sustainable solutions for water resources, as these solutions require systems based on machine learning and allow the collection and analysis of a huge amount of data to reach future visions.

When it is required to specify an output associated with different variable inputs, and the physical and chemical processes are not precisely and explicitly related, it is difficult to rely on mathematical modelling, since the relationships between the parameters are complex and non-linear (Kim and Parnichkun, 2017)

In the drinking water treatment process, reactions that are not well understood can frequently occur. This makes it very difficult to develop a useful mechanical model. Hence, applications of artificial intelligence of various kinds have been turned.

Many different recent studies have used different algorithms in modelling drinking water treatment plants; ANN and MLR models were used to predict the soluble sulphate content in drinking water (Tahraoui et al., 2021). MLR and ANN for predicting Residual aluminium (Tomperi et al., 2013). Hybrid ANN-GA and GEP for predicting Residual Turbidity (Alsaeed et al., 2021). GEP to predict turbidity (Wang et al., 2020). DT for modelling DOC (Tahraoui et al., 2022). GEP for predicting the Turbidity Removal using PACL (Alsaeed, 2021). ERT for predicting the coagulant dosage (Heddam and Dechemi, 2015). ANN was used to predict bicarbonate content of surface waters (Tahraoui et al., 2020). GEP for predicting the coagulant dosage (Alsaeed et al., 2022b).

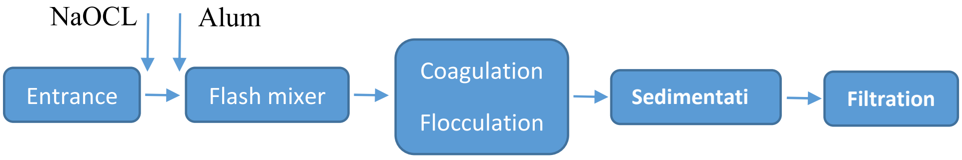
In the present research, three different methods: artificial neural networks, gene expressions and decision tree, were used to determine the values of residual aluminium in a drinking water plant. The models were built based on data form the plant for about four years, from a drinking water treatment plant; the coagulant used in this plant is alum. The models' results accuracy of the three techniques was compared.

2 Materials and methods

2.1 Models data

Research took place in Al Qusayr plant, Homs. Using a set of daily data, three different artificial intelligence methods were studied to determine the values of predicted aluminium in drinking water treatment plants. Figure 1 shows a schematic overview of the process at Qusayr WTP.

Figure 1 Al Qusayr plant process (see online version for colours)



The data used was a daily records from the plant; it is described in Table 1.

Table 1 Statistical characteristics of water samples

Parameter	*Turbidity* (Mg/L)	*Conductivity*	*pH*	*T* (°C)	*Alum dose* (Mg/L)	*Residual aluminium* (Mg/L)
Min	6.4	308	6.3	7.6	0	0
Max	65	420	9.3	22.6	24	0.33
Mean	19	347	7.43	15. 9	10	0.09
Std. Deviation	11	19.3	0.35	3.1	5.3	0.05

2.2 Data clustering

One of the most used algorithms is k-means clustering, which uses centroid-based approach to minimise intra-cluster variation. This method divides the total number of the data into k clusters. k-means clustering algorithm proceeds as follows:

$$j = \sum_{i=1}^{k} \sum_{xj \in Si} x_j - c_i^2 \tag{1}$$

where J = the objective function; x_j = the data vector given a set of observations (j = 1, 2, …, n), k = the number of clusters; Si is cluster; and c_i is cluster centre.

2.3 Neural network

Neural network is a formula that inherits human nerve cell capability. This capability allows it to perform prediction, classification, and pattern matching.

$$a_i = \sum_{j=1}^{n} x_i * w_i \quad +bi \tag{2}$$

A simple example of an artificial neural network:

- Input: $x_1 = 3, x_2 = 1, x_3 = 2$.
- Weighting coefficients: $w_1 = 3$, $w_2 = 0.4$, $w_3 = 0.4$.
- Summation function: $y = 3*0.2+1*0.4+2*0.4 = 1.8$
- Transformation function: $f(y) = \frac{1}{1+e^{-1.8}}$
- Output function: $Y = 0.85$.

Back progression is an algorithm that approaches the local minimum value of the error function by moving in steps proportional to the opposite direction of the error function gradient. We can define a function called the error function or the performance function to determine the difference between the actual output of the network and the desired output

$$E(W) = \frac{1}{2}\sum_{i=1}^{k}(a_a - a_p) \tag{3}$$

To reduce the error function, the weights are modified in the opposite direction of the gradient, i.e., in the direction:

$$D = -\nabla E(W_i) = -\frac{\partial E}{\partial W} \tag{4}$$

$$W_{i+1} = W_k - \mu.\nabla E(W_i) \tag{5}$$

2.4 Gene expression

It is one of the AI models; GEP is a type of genetic algorithm. It operates in the same way that a group abandons undesirable members and creates genetically engineered offspring in evolution.

GEP differs from standard GA in that it usually works with parse trees rather than bit strings. A terminal set (the problem's variables) and a function set are used to build a parse. Moreover, the GEP has in can solve problems in different fields with a high performance. Recently, this technique has been used to identify the behaviour of nonlinear systems.

The steps of GEP models could be summarised as the following: first the randomly creation the initial population generation. Then, the chromosomes are expressed and excluded the tree expression for fitness evaluating. The individual is then selected according to their fitness to reproduce with the modification; these individuals are subject to the same development. This process is going in repetition loop several times until a good solution is found.

2.5 Decision tree

Decision Tree Learning is a general, predictive modelling tool that has applications in different areas (Qin and Lawry, 2005).

Decision trees is one of the most widely used supervised learning methods. Decision Tree Learning is used for both classification and regression tasks. The decision rules are of the form ’if-then-else’. The deeper the tree, the more complex the rules and the model becomes better (Baldwin, Xie).

A decision tree is a tree-like graph with nodes representing the question, edges represent the answers to the question and the leaves represent the actual class label.

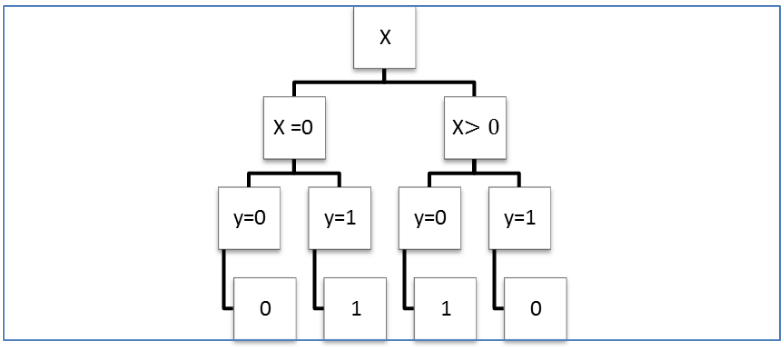
Compared to other data mining methods, the decision tree method is simpler to understand and interpret. It is easy to display graphically. It is suitable to handle both numerical and categorical data. And it performs well with large datasets.

Figure 2 Illustrates a simple decision tree model for the data listed in Table 2, for prediction M based on *X*, *Y*, *Z*.

Table 2 Example model data

X	*Y*	*Z*	*M*
0	0	0	0
0	1	1	1
0	0	1	0
0	1	0	1
1	0	1	1
1	1	0	0
1	0	0	1
1	1	1	0

Figure 2 Sample decision tree (see online version for colours)



Once a decision tree is built, it can be used to evaluate other samples and the results depends on how well it models the dataset.

The main components of a decision tree model are nodes, branches and the most important step is splitting.

Nodes: There are three types of nodes. A root node, or decision node, internal nodes, or chance nodes, and Leaf nodes, also called end node.

Branches: Branches represent chance outcomes that emanate from root nodes and internal nodes.

Splitting: to split parent nodes into purer child nodes of the target variable variables related to the target variable are used (Song and Ying, 2015).

2.6 Input selection

Determined are the relationships between the input variables and aluminium residual was defined by a statistical index called Pearson's correlation coefficient.

The correlation coefficient as 'shown in Table 3' of alum dose and the RA is the highest at 0.503 compared with the other input, thus is the most relevant to the output. Moreover, the input parameters are quite correlated among themselves as well.

Table 3 Pearson correlation coefficients of each input and aluminium

	Tur- in	*pH*	*T*	*Alum dose*	*Conductivity*	*AL*
Tur- in						
pH	–0.006					
T	–0.41	0.091				
Alum dose	0.71	–0.031	–0.5			
Conductivity	0.38	0.07	–0.48	0.38		
Al	0.32	0.21	–0.42	0.6	0.05	

Figure 3 shows the relation between many different parameters and the residual aluminium from the data taken from AlQusayer plant.

The possible inputs patterns are listed in the Table 4, the models without both raw turbidity and dose was not considered, as they are the most related parameters to the aluminium residuals, and can be used to make a model that could be generalised.

2.7 Models evaluation

The performance of various models was evaluated using the statistical indices:

- Root mean squared error (RMSE)

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\boldsymbol{T}_i - \boldsymbol{O}_i)^2}{\boldsymbol{n}}} \tag{6}$$

- Correlation coefficient (R) (HICHE)

$$\mathbf{R} = \frac{\sum_{i=1}^{n}\left(P_{obs} - \overline{\boldsymbol{P_{obs}}}\right)\left(P_{pre} - \overline{P_{pre}}\right)}{\sqrt{\sum_{i=1}^{n}\left(P_{obs} - \overline{\boldsymbol{P_{obs}}}\right)^2 \times \sum_{i=1}^{n}\left(P_{pre} - \overline{P_{pre}}\right)^2}} \tag{7}$$

- Mean absolute percentage error

$$MAPE = \frac{1}{n} * \sum_{t=1}^{n}\left|\frac{Y_{obs} - \hat{Y}_{\Pr e}}{n}\right| *100 \tag{8}$$

Y_{obs} : observed values , Y_{pre} : predicted values , $\overline{Y}_{obs}$: mean of observed values, $\overline{Y}_{pre}$: mean of predicted values.

Figure 3 Relation between different parameters and the residual aluminium (see online version for colours)

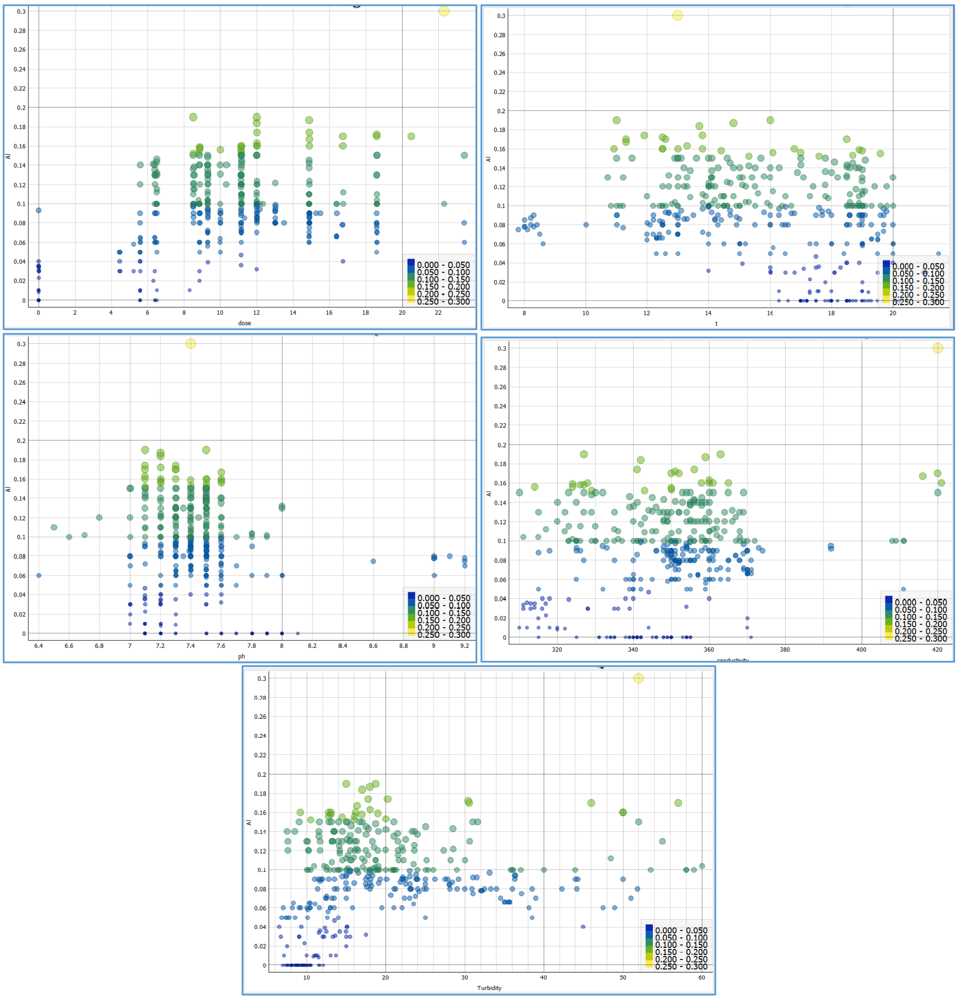


Table 4 Possible input combination for models

Input type	*Turbidity*	*Temperature*	*pH*	*Conductivity*	*Dose*
Model 1	I				I
Model 2	I	I			I
Model 3	I		I		I
Model 4	I			I	I
Model 5	I		I	I	I
Model 6	I	I		I	I
Model 7	I	I	I		I
Model 8	I	I	I	I	I

3 Results and discussion

The data had been processed and outliers were excluded. These values can hinder the proper training of the neural network and greatly affect its performance, so the entire record is excluded in case there were an extreme or missing value.

The data were clustered using K means algorithm, and the mean values of each parameters in the clusters is presented in Table 5.

Table 5 The mean values of the parameters in each cluster

Cluster number	*Parameter mean*				
	Dose	*T*	*Conductivity*	*pH*	*Turbidity*
1	8.6	17.2	322	7.2	14.3
2	19	8.5	350	9	33.5
3	12.3	12.3	362	7.5	22
4	6.5	19	349	7.5	11.3
5	15.1	16.5	355	7.1	46

A Q-Q plot of the parameters is shown in Figure 4.

The three sets of data was randomly taken from the five clusters. The statistical description of the datasets is described in Table 6.

The input and the output data obtained were normalised because they are of different ranges and units, otherwise there will be no correlation between the input and the output values. The data was normalised in the range [1,0].

Researchers have used different data division methods, There is no specific rule for data division, it varies according to problem type. In this study, adopted data division among training, validating, and testing sets is determined as (80%, 10%, 10%), respectively.

Figure 4 Q-Q plot for the parameters (see online version for colours)

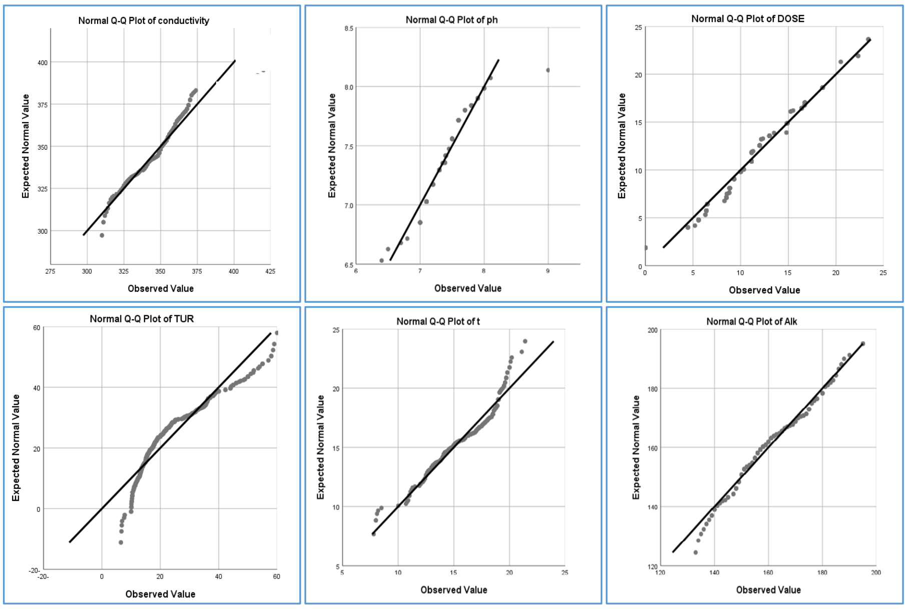


Table 6 Statistical properties for datasets

Datasets	*Parameters*	*min*	*max*	*Mean*
Training set	Turbidity	8.9	60	18
	Dose	0	24.1	12
	pH	7.1	8.8	7.3
	Temperature	8	23.4	15
	Conductivity	336	421	351
Validation set	Turbidity	7.9	32	12.5
	Dose	0	17.1	9.3
	pH	7	9.2	7.3
	Temperature	12.7	20	16.4
	Conductivity	311	416	348
Testing set	Turbidity	6.4	30	10
	Dose	0	15.1	8
	pH	6.8	9	7.5
	Temperature	7.6	17	17.4
	Conductivity	310	371	346

The architecture of the network was determined using Search Architect, it gave many architects of the network.

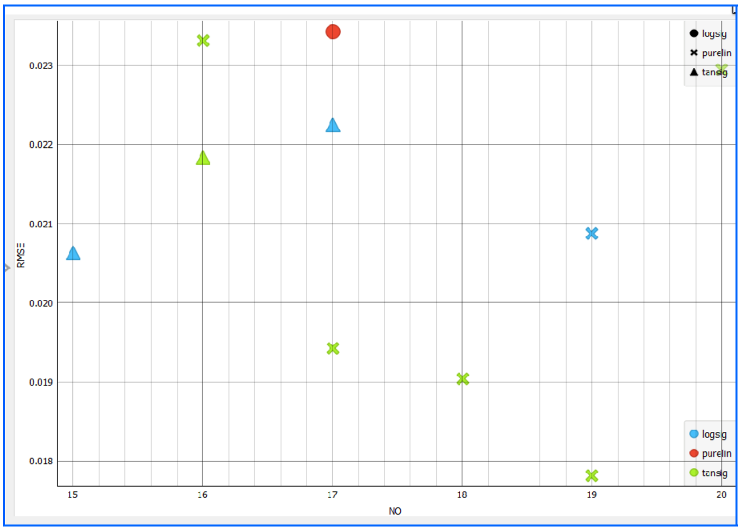
$$n_h \le \frac{n_T - n_0}{4\left(n_i + n_0 + 1\right)} \; \frac{300-1}{4*(5+1+1)} \le 11 \text{ (Loquasto and Seborg, 2003)} \tag{9}$$

$$n_h \leq \sqrt{n_i \;.n_T} \;\; \leq \sqrt{5\;.300} \;\; = 39 \text{ (Verma, 2014)} \tag{10}$$

n_T : number of training data, n_0 : number of outputs, n_i : number of inputs.

The best 10 networks' results are shown in Figure 5.

Figure 5 Comparing the results of the top 10 resulting networks (see online version for colours)



From the previous results, it can be concluded that the best architecture for the neural network was (5-19-1); it means 5 inputs, 19 neurons, and one output. It was trained with LM algorithm. It gave a good predicting ability; the results were good and the network was able to predict the residual aluminium. The model could predict the minimum and maximum values with a good accuracy, the results are shown in Table 7.

Table 7 The network ability of prediction

	Training	*Validation*	*Testing*
RMSE	0.0175	0.018	0.022
R	0.96	0.93	0.9

3.1 *Gene expression model*

First, 75% of the data available from the drinking water purification plant was used to train the model. The remaining 25% of the data were used for validation. Fitness function was RMSE, and a set of functions was selected.

There is a variety of parameters related to the gene expression models; the most important is the number of chromosomes, mutation and the set of functions.

In this study five different numbers of chromosomes were used.

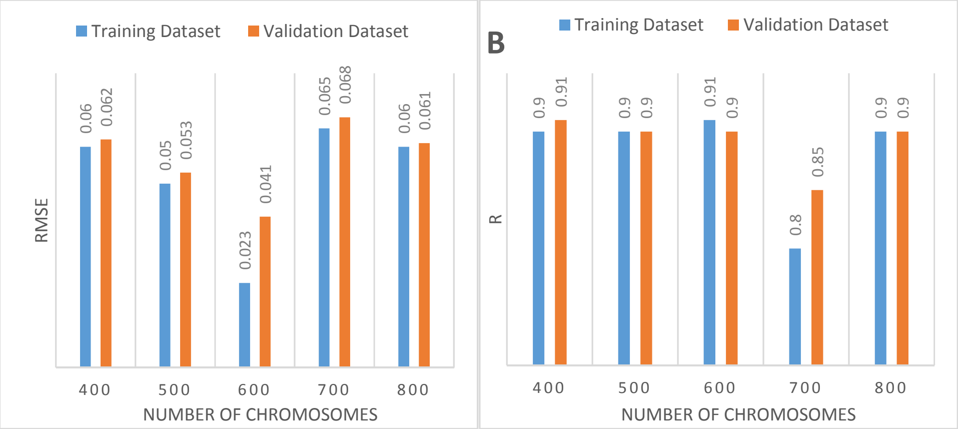
The values of GEP parameters are shown in Table 8.

Table 8 Values of GEP control parameters

Function set	+, –, ×, /, √, *exp, ln, 10^*
Constants per gene	10
Function Fitness	RMSE
Linking Function	+
Mutation	0.00138
Head size	9

Different numbers of chromosomes (all with the same parameters listed in Table 8) were tested, and the results are listed in Figure 6.

Figure 6 Comparison between (A- RMSE) and (B- R) of training and validation for a different numbers of chromosomes (see online version for colours)



The results for the best two models are shown in Table 9.

Two gene-expression trees were developed for the best model, the sub gene-expression trees are described in Figure 7.

Table 9 Results of GEP best two models

The number of the model	*Inputs used*		*RMSE*	*R*
1	Turbidity, Temperature, Alum Dose , pH, conductivity	Train	0.023	0.91
		Test	0.027	0.9
2	Turbidity, Temperature, Alum Dose , pH	Train	0.028	0.82
		Test	0.035	0.81

RMSE in the training was higher than in the testing phase.

The constants and the parameters used in the equation shown in Figure 6 are listed in Table 10.

Figure 7 The sub gene-expression tree (see online version for colours)

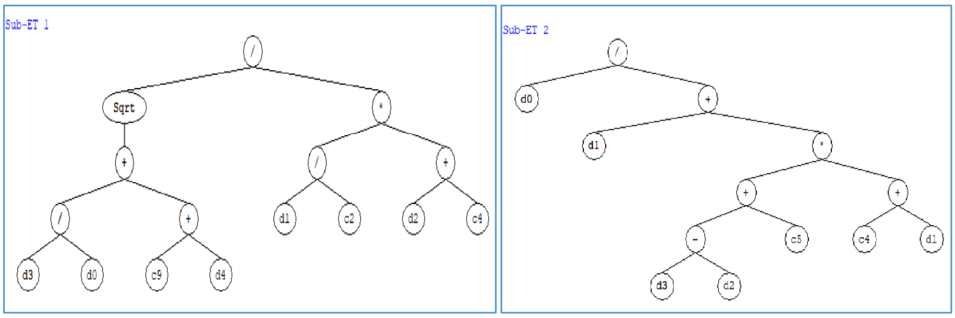


Table 10 Values of parameters used in in the equations

d_0	d_1	d_2	d_3	d_4
Turbidity	Conductivity	T	pH	Alum dose
	G1C2		49.4535194770961	
	G1C4		–4.28765040286449	
	G1C9		–1.08676759970002	
	G2C5		–4.55590869124656	
	G2C4		–101.355192214958	

3.2 Decision tree

When developing the model, the most relevant input variables should be determined, and records should be separated into two or more categories based on the status of these variables at the root node and subsequent internal nodes.

This process of splitting continues until the homogeneity or halting requirements are reached. In most circumstances, not all possible input variables will be used to construct the decision tree model, and a single input variable may be used numerous times at different levels of the decision tree. Number of instance in leaf = 3.

The resulted tree, described in Figure 8, could predict aluminium with an acceptable results, RMSE = 0.027 mg/L and $R = 0.91$. DT was the least good model in predicting the aluminium.

Figure 8 The subtree *T* (see online version for colours)

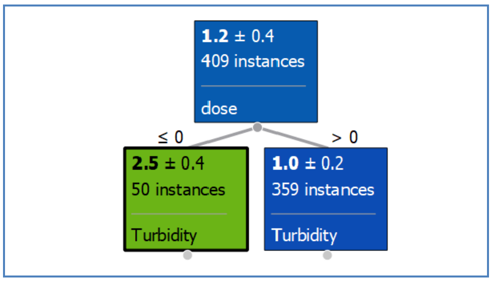


Figure 8(a) The subtree T_1 of T (see online version for colours)

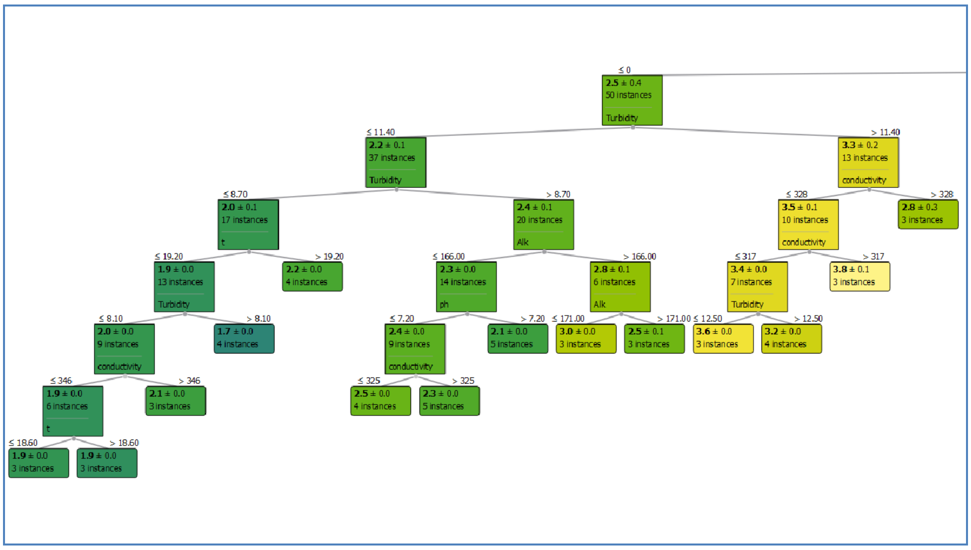


Figure 8(b) The subtree T_2 of T (see online version for colours)

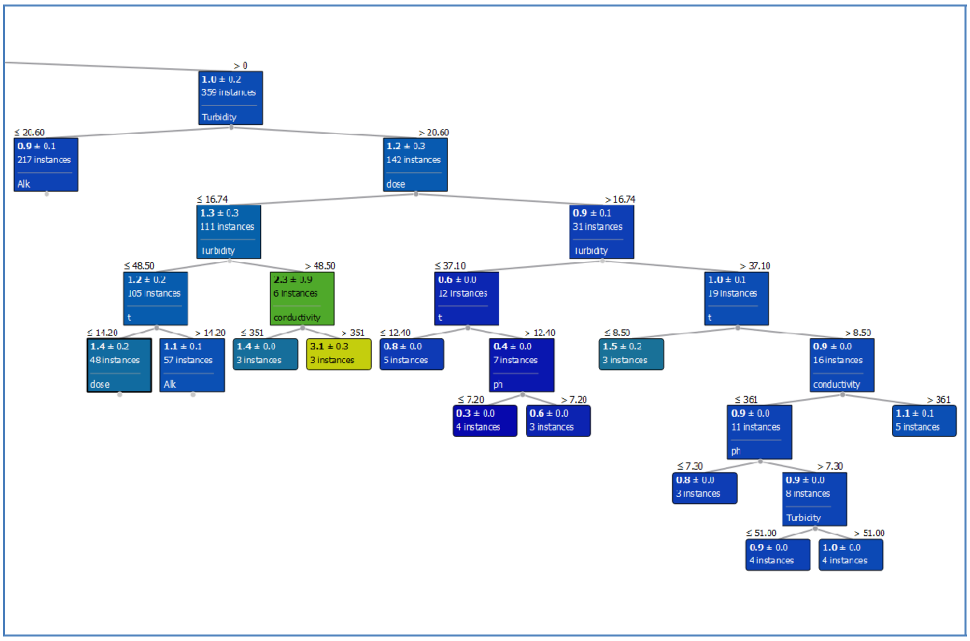
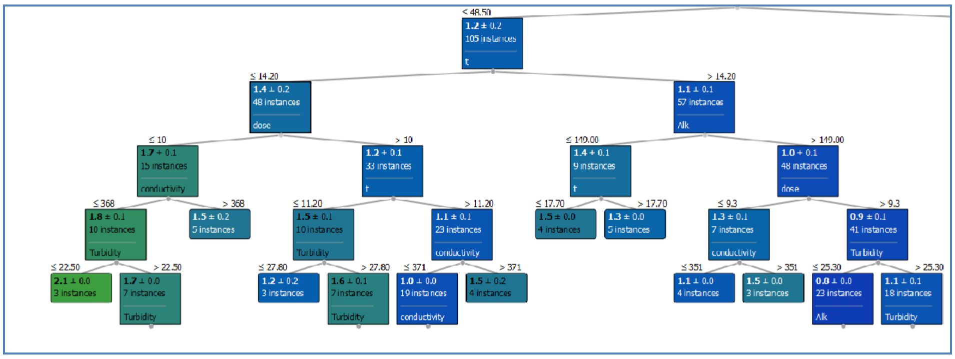


Figure 8(c) The subtree T_3 of T (see online version for colours)



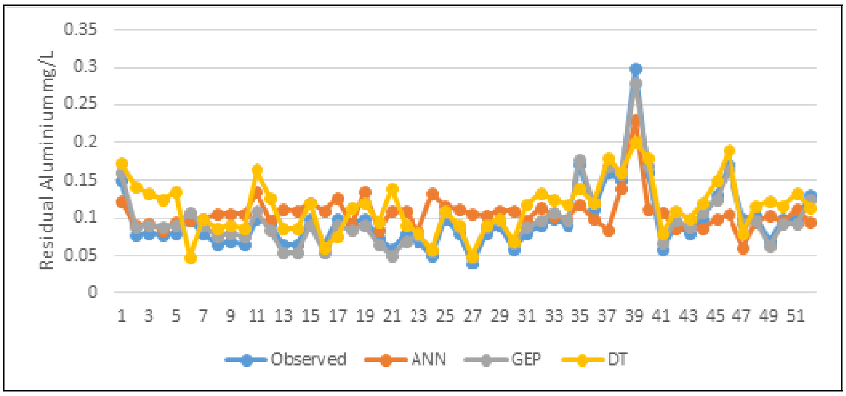
When comparing the results, it was found that, for initial turbidity between 6.5 NTU and 30 NTU, ANN provides the best performance. For turbidity from 30 NTU to 60 NTU which are the prevailing qualities of raw water, GEP is better than ANN and DT models. These results are listed in Table 11.

Table 11 Comparing ANN, GEP, and DT by initial turbidity

Statistical measurements of turbidity	*RMSE*			*Number of data*
	ANN	*GEP*	*DT*	
6.5 < Turbidity < 30	**0.016**	0.026	0.028	259
30 < Turbidity < 60	0.024	**0.023**	0.025	41
Entire data	**0.018**	0.025	0.027	300

Figure 9 represents a comparing chart of part of the data with the results gained from the three models, for turbidity higher than 30 as it is shown the GEP model was ahead in this range of initial turbidity.

Figure 9 The observed and modelled values (see online version for colours)



4 Conclusion

In this research, three models were built to predict the values of the residual aluminium in water treatment plants, as it gives signs that the coagulation process has been in its optimal form. The three used algorithms gave good results in the simulation process, ANN was slightly better than the other two models, with RMSE = 0.019 mg/L and R = 0.94. GEP gave also good results, GEP equation is able to be easily used, without the need of a codes make and this makes it the most easily generalised model.

Data availability

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of interest

The authors declare no conflict of interest.

References

Alsaeed, R. (2021) 'Modelling turbidity removal by poly-aluminium chloride coagulant using gene expression', *Advances in Environmental Technology*, Vol. 7, No. 4, pp.263–273.

Alsaeed, R., Alaji, B. and Ebrahim, M. (2021) 'Predicting turbidity and aluminum in drinking water treatment plants using hybrid network (GA-ANN) GEP drink', *Water Eng. Sci. Discuss*, [preprint], https://doi.org/10.5194/dwes-2021-8

Alsaeed, R.D., Alaji, B. and Ibrahim, M. (2022a) 'Modeling jar test results using gene expression to determine the optimal alum dose in drinking water treatment plants', *Baghdad Science Journal, Iraq, Baghdad*, Vol. 19, No. 5, p.0951, doi: 10.21123/bsj.2022.6452.

Alsaeed, R.D., Alaji, B. and Ibrahim, M. (2022b) 'Using bentonite clay as coagulant aid for removing low to medium turbidity levels', *Iranian Journal of Chemistry and Chemical Engineering*, Vol. 41, No. 12, https://www.ijcce.ac.ir/article_249404.html

Amin, D. and Sadaf, H. (2018) 'Optimum coagulant forecasting by modeling jar test experiments using ANNs', *Journal of Drinking Engineering and Science*, doi.org/10.5194/dwes-11-1-2018.

Heddam, S. and Dechemi, N. (2015) 'A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modelling coagulant dosage (Dos): case study of water treatment plant of Algeria', *Desalination and Water Treatment*, Vol. 53, No. 4, pp.1045–1053.

Kim, C.M. and Parnichkun, M. (2017) 'Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system', *Applied Water Science*, Vol. 7, pp.3885–3902.

Kim, S. and Yoon, C. (2000) *Reducing Residual Alum Concertation at Water Treatment Plant by Improving Filtration Performances*, Kuyngnum University, Korea.

Krupińska, I. (2020) 'Aluminium drinking water treatment residuals and their toxic impact on human health', *Molecules*, Vol. 25, No. 3, p.641.

Loquasto, F. and Seborg, D.E. (2003) 'Monitoring model predictive control systems using pattern classification and neural networks', *Industrial & Engineering Chemistry Research*, Vol. 42, No. 20, pp.4689–4701.

Qin, Z. and Lawry, J. (2005) 'Decision tree learning with fuzzy labels', *Information Sciences*, Vol. 172, Nos. 1–2, pp.91–129.

Song, Y.Y. and Ying, L.U. (2015) 'Decision tree methods: applications for classification and prediction', *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, p.130.

Tahraoui, H., Amrane, A., Belhadj, A.E. and Zhang, J. (2022) 'Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting', *Environmental Technology and Innovation*, Vol. 27, p.102419.

Tahraoui, H., Belhadj, A.E. and Hamitouche, A.E. (2020) 'Prediction of the bicarbonate amount in drinking water in the region of médéa using artificial neural network modelling', *Kemija u Industriji: Časopis Kemičara i Kemijskih inženjera Hrvatske*, Vol. 9, Nos. 11–12, pp.595–602.

Tahraoui, H., Belhadj, A.E., Hamitouche, A.E., Bouhedda, M. and Amrane, A. (2021) 'Predicting the concentration of sulfate (So4 2–) 'in drinking water using artificial neural networks: a case study: médéa-algeria', *Desalination and Water Treatment*, Vol. 217, pp.181–194.

Tomperi, J., Pelo, M. and Leivisk, K. (2013) *Predicting the Residual Aluminum Level in Water Treatment Process*, University of Oulu, Finland, doi. org/10.5194/dwes-6-39-2013.

Verma, A.K. (2014) *Process Modelling and Simulation in Chemical, Biochemical and Environmental Engineering*, CRC Press.

Wang, Y., Feng, L., Liu, J., Hou, X. and Chen, D. (2020) 'Changes of inundation area and water turbidity of Tonle Sap Lake: responses to climate changes or upstream dam construction?', *Environmental Research Letters*, Vol. 15, No. 9, p.0940a1.