Drinking Water Assessment Using Statistical Analyses of AL-Muthana Water Treatment Plant

Mohammed Abed Naser[1,2(✉)] and Khalid Adel Abdulrazzaq[1]

[1] Civil Engineering Department, University of Baghdad, Baghdad, Iraq
mohammadnasier0@gmail.com, aleoubaidy@coeng.uobaghdad.edu.iq
[2] Directorate of Education Al-Muthana, Ministry of Education, Baghdad, Iraq

Abstract. Water is essential for survival, and controlling water quality is one of the most basic requirements for protecting this natural wealth from pollution and extinction. Statistical analysis technique was used with the SPSS v26 software to evaluate the quality of raw water period from (2016–2020), and 14 water parameters were assessed (Ka, Na, TSS, TDS, Mg, Ca, SO_4, Alk, TH, pH, Turbid, Temp, Cl, and Ec). Five principal components have eigenvalues value greater than unity and explain (76.159%) of the total variance of original data set. The first component was (28.678%) of the total variance with high loading on (TH, Ca, Mg, Cl and Ka), the second component was (16.141%) with positive loading on (TSS, Turb, and Temperature), the third component was (14.826%) with positive loading on (TDS and Ec), the fourth component was (8.929%) with positive loading in (Alk and SO_4), and the last one has (7.59%) from total variance which high positive loading in (pH). The Multiple Linear Regression (MLR) results in a strong relationship between water conductivity and total suspended solid with other water parameters which the coefficient of determination (R^2) values were 0.963 and 0.92. The ANN model was created to forecast river turbidity based on influent TSS, Mg, TDS, and Ca. The sum of squared errors and relative errors being (0.231, 0.101) and (0.009, 0.027) respectively, respectively, the error rate in predicting the model is low, indicating that the model is successful in predicting the turbidity of the river's raw water.

Keywords: Drinking water · Assessment · Statistical analyses

1 Introduction

Water is one of the most important basic sources of life. Water pollution has a significant impact on human health, so keeping clean water free of pollution is a high priority for a disease-free life [1]. Detecting variations in the quality of drinking water through the use of statistical analysis techniques, which improve the reliability of the system in laboratories, allowing the administration to make the appropriate decision, and these techniques provide foresight into future changes and challenges confronting the authorities in charge of managing and regulating water [2]. Monitoring water quality over time and using statistical analysis is one of the most common and effective methods for evaluating time changes and environmental problems that occur in raw water sources

M. Karkush et al. (Eds.): ICGECI 2022, *Current Trends in Geotechnical Engineering and Construction*, pp. 13–20, 2023.
https://doi.org/10.1007/978-981-19-7358-1_2

based on chemical and physical parameters and biological indicators, and it contributes significantly to assisting researchers in changing the state of pollution [3]. According to [4], the temporal data of raw water quality is a fundamental method for discovering hypotheses that were not present when measurements were taken and were not expected, and it is of a standard value that reveals important patterns that allow us to identify trends and rare events that appear, and thus discover undesirable characteristics in the quality of drinking water. One of the most important components of machine learning and artificial intelligence is artificial neural networks. It is inspired by the structure of the human brain and operates as if it were made up of interconnected nodes where simple manipulations can be performed [5].

Artificial Neural Networks (ANNs), also known as neural networks, are cutting-edge computational systems and methods for deep learning, knowledge presentation, and finally applying acquired knowledge to maximize complicated system output responses [6]. Also, the principal component analysis (PCA) is another statistic technique used in water quality sample testing to summarize and abbreviate data, converting a large number of implicitly, albeit partially, correlated variables into a much smaller set of imaginary independent variables, which are usually called principal components. It is calculated primarily from the original variables in ratios and amounts that increase or decrease depending on the role and influence of the original variables [7]. This study aims to use statistical analysis tools to assess drinking water quality in Al-Muthanna Water Treatment Plant and find mathematical models that allow us to predict basic water quality parameters.

2 Materials and Methods

The following points can summarize the statistical analysis and modeling prediction of source water using SPSS v26:

- Multiple linear regression (stepwise regression model) was used for raw water quality parameters to find a mathematical relationship that predicts the value of (Ec and Turb) with other water parameters.
- Principal component analysis (PCA): is a technique for identifying a smaller number of uncorrelated variables known as principal components from a larger set of data. This technique is commonly used to highlight differences and capture strong patterns in data sets. It is one of the most widely used methods for analyzing water quality data and reducing variables without affecting the system [8].
- The ANN model was created to forecast river turbidity based on influent TSS, Mg, TDS, and Ca turbidity. Water turbidity is one of the most important parameters indicating the quality of drinking water because it is an integrated parameter that is closely related to the rest of the water quality parameters and can be used to infer the quality of drinking water [9].

3 Results and Discussion

3.1 Principal Component Analysis (PCA)

The data were standardized, and the Kaiser-Meyer-Olkin (KMO) and Bartlett sphericity tests were computed. The KMO test resulted in 0.62, and the Bartlett sphericity test resulted in less than 0.001. The value of (Kaiser-Meyer-Olkin (KMO) test and Bartlett sphericity test) equals (0.66), which is an acceptable value because the minimum value is (0.5), indicating that the measurement is good and the significant degree of the measurement has been reached (0). Eigenvalues accounts and scree plot in Table 1 listed that only five principal components have eigenvalues greater than unity and explain (76.159%) of the total variance in the data set. The first component accounted for about (28.678%) of the total variance with high loading on (TH, Ca, Mg, Cl, and Ka). Polyvalent mineral ions cause water hardness. Water hardness is formed primarily by calcium and magnesium. An increase in hardness has a negative impact on human health and the industries that use water [10]. The second component accounted for about (16.141%) of the total variance and has a height positive loading for (TSS, Turb, and Temperature). The increase in suspended matter and turbidity has a negative impact on drinking water quality and the efficiency of drinking water treatment plants [11].

The third component accounted for about (14.826%) of the total variance and has a height positive loading for (TDS and Ec). TDS and Ec are water quality parameters that indicate salinity. These two parameters are correlated and are usually expressed using the following simple equation: k Ec = TDS [12]. The fourth component accounted for about (8.929%) of the total variance and has a height positive loading in (Alk and SO_4), and the last one has (7.59%) from total variance which high positive loading in (pH). Increased pH and Alk values indicate that the water contains alkaline salts like (NaOH) and $(CaOH)_2$ [13].

Table 1. Total variance explained.

Comp.	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total
1	4.014	28.672	28.672	4.014	28.672	28.672	3.383
2	2.260	16.141	44.813	2.260	16.141	44.813	3.246
3	2.076	14.826	59.640	2.076	14.826	59.640	2.304
4	1.250	8.929	68.569	1.250	8.929	68.569	1.721
5	1.063	7.590	76.159	1.063	7.590	76.159	1.268

3.2 Multiple Regression Results (MLR)

Multiple linear regression is an advanced statistical method that ensures the accuracy of inference in order to improve research results through the optimal use of data in finding causal relationships between the phenomena of the subject of study. The multi-linear regression analysis was used to determine the relationship between the dependent and independent water quality parameters. The dependent variables (TSS and Ec) were selected, and the most independent influence parameter of water quality was chosen using a stepwise regression mod. The Multiple Regression Results listed in Table 2 and Fig. 1 a strong relationship between water conductivity and other water parameters, with a coefficient of determination (R^2) value (0.963), indicating a positive relationship with a high degree of predictability. The test also indicated that electrical conductivity is a function of (TDS) which can be represented in Eq. (1) that can be used to predict current and future values

$$\text{Ec} = 55.917 + 1.595\ \text{TDS} \tag{1}$$

Table 2. Model summary of Ec prediction**.**

Model	R	R^2	Adjusted R^2	Std. error of the estimate
1	.981[a]	.963	.953	10.4614

[a]Dependent variable Ec

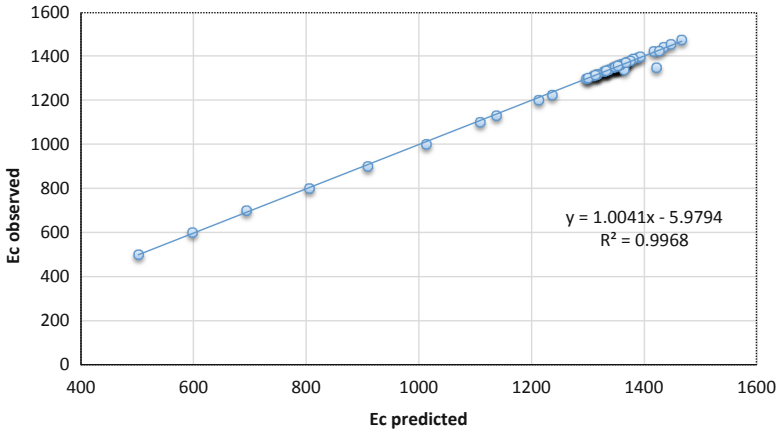

Fig. 1. Comparison of Ec (μS/cm) between observed and predicted values.

Table 3 and Fig. 2 listed that high relegation factor between (TSS) and other water quality parameters with (R^2) value (0.92), and indicated that total suspended solid is a function of (Turbidity), which can be produced by the Eq. (2).

$$\text{TSS} = 0.93 + 1.903\ \text{Turb} \tag{2}$$

Table 3. Model summary of TSS prediction**.**

Model	R	R^2	Adjusted R^2	Std. Error of the estimate
1	.959[a]	.920	.897	6.5505

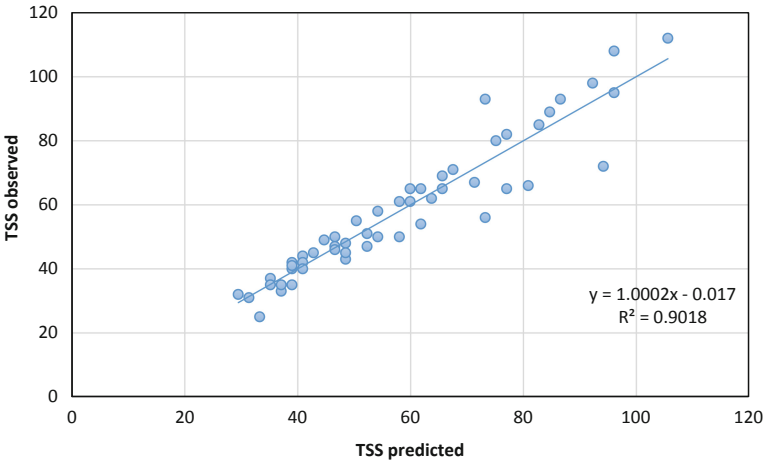

Fig. 2. Comparison of TSS (mg/L) between observed and predicted values.

Both above equations have been tested in the field and have proven to be reliable equations capable of producing very high results.

3.3 Artificial Neural Network (ANN)

The neural network application is an important tool for predicting water quality in drinking water treatment plants in order to reduce analysis and operating costs, evaluate performance, and control operating conditions more broadly [14]. The ANN model was created to forecast river turbidity based on influent TSS, Mg, TDS, and Ca. The model required 61 input data points divided into 47 for training and 14 for testing. Standardization is the data scaling method, and the number of hidden layers is one. Table 4 listed the amount of error was small in both training and testing, with the sum of squared errors and relative errors being (0.231 and 0.101) and (0.009 and 0.027) respectively, which the error rate in predicting the model is low, indicating that the model is successful in predicting the turbidity of the river's raw water. Table 5 explains the nature of the strength of the relationship between the independent and dependent factors in the input, hidden, and output layers, illustrated in Fig. 3. It shows that the gray line indicates positive values, and the blue line indicates negative values. The thickness of the line depends on the element's influence in the prediction process, regardless of whether the value is positive or negative.

Table 4. ANN model summary.

Training	Sum of squares error	0.213
	Relative error	0.009
	Stopping rule used	1 consecutive step(s) with no decrease in error
	Training time	0:00:00.00
Testing	Sum of squares error	0.101
	Relative error	0.027
Dependent Variable: Turbidity		
a. Error computations are based on the testing sample		

Table 5. Hidden layer parameters parameter estimates.

Predictor		Predicted			
		Hidden layer 1			Output layer
		H (1:1)	H (1:2)	H (1:3)	Turbidity
Input layer	(Bias)	.118	.098	−.246	
	TSS	−.405	−.541	.420	
	Ca	−.026	−.122	−.337	
	TDS	.025	−.470	−.142	
	Mg	.169	−.307	.305	
Hidden layer 1	(Bias)				.332
	H (1:1)				−1.624
	H (1:2)				−.355
	H (1:3)				.611

Figure 4 compares the prediction of turbidity concentrations based on different turbidity input parameters observed in this study. The expected trend follows the observed trend for all input data, and there is a significant convergence between the expected and actual values, with a small disparity.

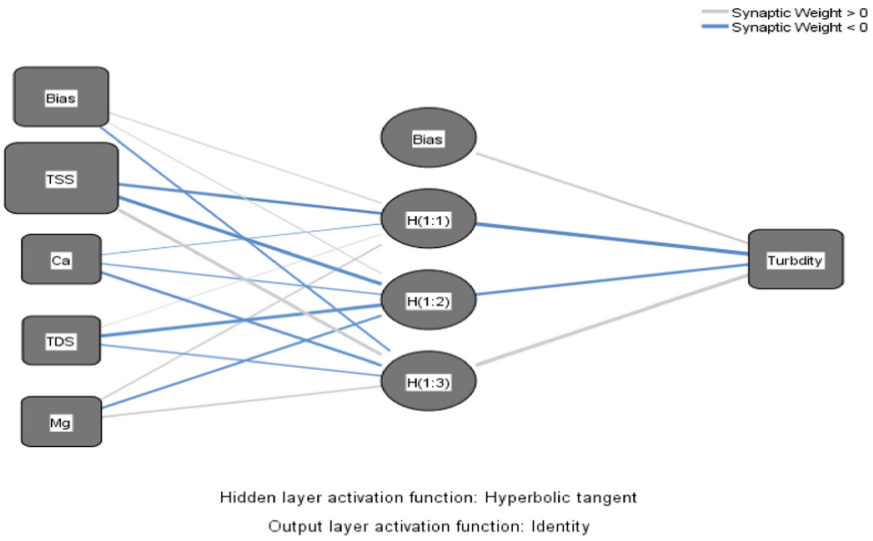

Fig. 3. Artificial neural network.

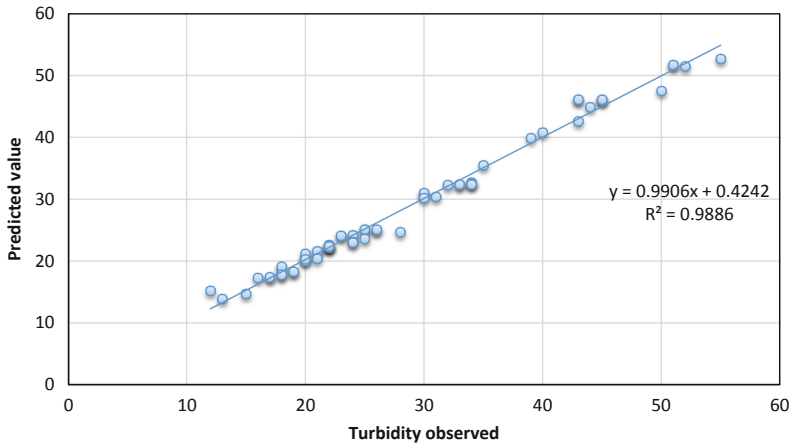

Fig. 4. Actual turbidity (NTU) versus predicted turbidity.

4 Conclusions

The following were the conclusions reached through the use of statistical analysis techniques: -

- Only five principal components have eigenvalues value greater than unity and explain (76.159%) of the total variance in the data set.

- The Multiple Linear Regression (MLR) results in a strong relationship between water conductivity and total suspended solid with other water parameters which the coefficient of determination (R^2) values were (0.963) and (0.92) sequentially, indicating a positive relationship with a high degree of predictability.
- The ANN modeling results show that the model performed at 99.06% prediction accuracy

References

1. Juntunen, P., Liukkonen, M., Pelo, M., Lehtola, M.J., Hiltunen, Y.: Modelling of water quality: an application to a water treatment process. Appl. Comput. Intell. Soft Comput. **2012** (2012)
2. Nnorom, I.C., Ewuzie, U., Eze, S.O.: Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria. Heliyon **5**(1), e01123 (2019)
3. Sun, X., et al.: Analyses on the temporal and spatial characteristics of water quality in a seagoing river using multivariate statistical techniques: a case study in the Duliujian River, China. Int. J. Environ. Res. Public Health **16**(6), 1020 (2019)
4. Burt, T.P., Howden, N.J.K., Worrall, F.: On the importance of very long-term water quality records. Wiley Interdiscip. Rev. Water **1**(1), 41–48 (2014)
5. Kujawa, S., Niedbała, G.: Artificial neural networks in agriculture. Agriculture **11**(6), 497 (2021)
6. Chen, M., Challita, U., Saad, W., Yin, C., Debbah, M.: Artificial neural networks-based machine learning for wireless networks: a tutorial. IEEE Commun. Surv. Tutor. **21**(4), 3039–3071 (2019)
7. Teixeira de Souza, A., Carneiro, L.A.T., da Silva Junior, O.P., de Carvalho, S.L., Américo-Pinheiro, J.H.P.: Assessment of water quality using principal component analysis: a case study of the Marrecas stream basin in Brazil. Environ. Technol. **42**(27), 4286–4295 (2021)
8. Zavareh, M., Maggioni, V., Sokolov, V.: Investigating water quality data using principal component analysis and granger causality. Water **13**(3), 343 (2021)
9. Iglesias, C., et al.: Turbidity prediction in a river basin by using artificial neural networks: a case study in northern Spain. Water Resour. Manage **28**(2), 319–331 (2014)
10. Akram, S., Rehman, F.: Hardness in drinking-water, its sources, its effects on humans and its household treatment. J. Chem. Appl. **4**(1), 1–4 (2018)
11. Serajuddin, M., Chowdhury, A.I., Haque, M.M., Haque, M.E.: Using turbidity to determine total suspended solids in an urban stream: a case study. In: Proceedings of the 2nd International Conference on Water and Environmental Engineering, Dhaka, pp. 19–22 (2019)
12. Rusydi, A.F.: Correlation between conductivity and total dissolved solid in various type of water: a review. In: IOP Conference Series: Earth and Environmental Science, vol. 118, no. 1, p. 012019. IOP Publishing, February 2018
13. Putro, P.G.L., Hadiyanto, H.: Water quality parameters of tofu wastewater: a review. In: IOP Conference Series: Materials Science and Engineering, vol. 1156, no. 1, p. 012018. IOP Publishing, June 2021
14. Nasr, M.S., Moustafa, M.A., Seif, H.A., El Kobrosy, G.: Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. Alex. Eng. J. **51**(1), 37–43 (2012)