Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.26813

217 (2021) 181–194
March

Predicting the concentration of sulfate (SO_4^{2-}) in drinking water using artificial neural networks: a case study: Médéa-Algeria

Hichem Tahraoui[a], Abd-Elmouneïm Belhadj[a], Adhya-eddine Hamitouche[b], Mounir Bouhedda[c], Abdeltif Amrane[d,*]

[a]*Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Médéa, 26000 Médéa, Algeria, emails: tahraoui.hichem@univ-Médéa.dz (H. Tahraoui), belhadj_1@yahoo.fr (A.-E. Belhadj)*
[b]*Centre de Recherche scientifique et Technique en Analyses Physico-Chimiques CRAPC, BP 384, Bou-Ismail, RP 42004, Tipaza, Algeria, email: ahamitouche2@yahoo.fr*
[c]*Laboratory of Advanced Electronic Systems (LSEA), University of Médéa, 26000 Médéa, Algeria, email: bouhedda.m@gmail.com*
[d]*Univ Rennes, Ecole Nationale Supérieure de Chimie de Rennes, CNRS, ISCR – UMR6226, F-35000 Rennes, France, email: abdeltif.amrane@univ-rennes1.fr*

Received 19 April 2020; Accepted 12 November 2020

ABSTRACT

The aim of this work was to use artificial neural networks (ANN) and multiple linear regressions (MLR) models to predict the soluble sulfate content in drinking water. A set of 84 data points were used. For the ANN, 18 neurons were used in the input layer, 8 neurons at hidden layer, and 1 was used in the output layer. Levenberg Marquardt learning (LM) algorithm with hyperbolic tangent sigmoid transfer function logarithmic was used at the hidden and output layer. The comparison of the obtained results in term of root mean square error (RMSE) and correlation coefficient (R) using the ANN and MLR models revealed the superiority of the (ANN) model in predicting the soluble sulfate content in drinking water. Indeed, the statistical results showed a correlation coefficient R = 0.99973 with RMSE = 5.9755 for the ANN model and R = 0.941 with RMSE = 88.3068 for the MLR model. A nonlinear relationship between the soluble sulfate content and the physico-chemical characteristics of drinking water (conductivity, turbidity, potential hydrogen, hardness, calcium, magnesium, chlorides, total alkali metric titre, material organic, nitrogen dioxide, nitrates, sodium, bicarbonate, potassium, heavy metals (Mn^{2+}, Fe^{3+}, and Al^{+}) and dry residues) was demonstrated, showing that the soluble sulfate content concentration can be predicted.

Keywords: Drinking water; Physico-chemical parameters; Sulfate; Modeling; Artificial neural networks; Multiple linear regressions

1. Introduction

Water is a basic element for the sustainable development of the city. Drinking water is essential for human survival and regional stability [1]. Thus, water quality is an extremely important environmental factor as it affects human beings and their economic activities [2]. The main factor that affects the physical appearance such as the color of the water is the concentration and distribution of suspended fine components and dissolved matter [2]. Industrial, agricultural, and urban development is altering water quality and making it unsafe. This is the case of the Medea region, which is subject to various types of pollution and an increase in the quantity of pollutants

* Corresponding author.

released into the aquatic environment without treatment. Depending on the origin of the waste, the pollution can be of a chemical nature. Water quality tends to degenerate progressively with human interventions, such as hydrological alterations [3], land-use changes [4], inputs of toxic chemicals, and nutrients [5] and changes in other physico-chemical properties of water [6] this causes a range of environmental problems, such as the soluble sulfate content in drinking water levels. The soluble sulfate content in drinking water is ubiquitous and can be found not only in natural waters, but also in industrial wastewater [7,8]. Although the soluble sulfate content in drinking water is generally not considered a health problem, concentrations of soluble sulfate content in drinking water can cause a bitter taste and can cause diarrhea when its concentration exceeds 600 mg/L [9]. The release of high levels of sulfate can significantly affect the water supply by causing corrosion and/or scaling of pipes and equipment. In addition, hydrogen sulfide (H_2S), which is toxic to the ecosystem, could be produced by sulfate reduction by sulfate-reducing bacteria under anaerobic conditions [10]. Because of these adverse effects on human health and the environment, many countries have set maximum soluble sulfate content in drinking water concentration values ranging from 250 to 500 mg/L, depending on the end-use of the water source [8]. Traditional methods of measuring and studying water quality are both time-consuming and costly compared to numerical modeling techniques, especially if they deal with large areas. Recently, many authors have studied parameters affecting water quality using artificial intelligence. Several works have found great success in the simulation and prediction of environmental parameters [11] such as the prediction of several environmental parameters dealing with water quality in rivers of different countries [12–18], the prediction of the indicators of quality water for urban source management [19], analysis of surface water quality and identification of key water parameters [20], prediction of water quality via *Escherichia coli* levels [21], prediction of soil hard-setting, and physical quality using water retention data [22] and the prediction of phosphorus and total nitrogen for lake in Egypt [23]. In particular, artificial neural networks (ANNs) are a method for approximating complex systems, especially useful when these systems are difficult to model using classical statistical methods [24]. ANNs provide interesting results due to their learning capability [25], their parallelism, and their ability to solve many non-linear system problems [26]. Over the last decade, ANN research has been applied in the fields of hydrology, ecology, and the environment. ANN models have been shown to perform better than other models, the prediction of water quality [27]. ANN models have been requested for a variety of purposes; for example, for variations in water quality attributes [28], for prediction of water quality parameters [29–33], for prediction of water quality indices [34], for estimation of lake water quality using satellite images [35], to study water quality parameters of the Axios River in Northern Greece [36], to model nitrate concentrations in rivers [37], prediction of annual drinking water quality reduction based on groundwater resource index [38], prediction of the groundwater remediation costs for drinking use based on the quality of water resources [39]. In addition, the evaluation of multivariate linear regression and ANNs for predicting water quality parameters were examined [40].

The main objective of this research was therefore to conduct a comparative study between multiple linear regression (MLR) and ANN for the prediction of soluble sulfate content in drinking water in the Médéa region in Algeria.

2. Materials and methods

2.1. Database

The data used for this study were obtained from the experimental analysis of water samples taken during several sampling times during the period 2018 in the region of Médéa, Algeria (three samples per week from different regions). Analyses were done according to Jean Rodier's book of water analysis 9th edition [41].

The dependent variable was the soluble sulfate content in drinking water. The independent variables were the physicochemical parameters: conductivity, turbidity, potential hydrogen, hardness, calcium, magnesium, chlorides, total alkali metric titre (TAC), organic material, nitrogen dioxide, nitrates, sodium, bicarbonate, potassium, heavy metals (Mn^{2+}, Fe^{3+}, and Al^{+}), and dry residues.

2.2. Prediction methods

Several methods were applied to address problems related to prediction and modeling of complex nonlinear systems. These methods are particularly useful when these systems are difficult to model using classical methods [42]. In this study, we were interested in the use and the comparison of two methods for predicting the soluble sulfate content levels from the physicochemical parameters in drinking water. These methods are MLR and ANNs.

The data are divided into three phases (70% learning, 15% testing, and 15% validation) according to the ANN model to compare the correlation coefficients and the errors between the two models [43].

2.2.1. Multiple linear regressions

MLR consists in describing the relationships between a dependent variable y and several variables called independent variables $x_1,x_2,\ldots,x_i,\ldots,x_n$, where n is the number of independent variables. Indeed, the MLR, which is a data analysis method, is commonly used for establishing predictive models to the phenomena observed in the aquatic environment [44]. This method allows to draw a polynomial function describing the relationship between the dependant and the independent variables and allows to determine the most significant input variables. The model can be written as in Eq. (1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_n x_n + \varepsilon(x) \quad (1)$$

where n is the number of independent variables; $\varepsilon(x)$: random noise (error term or regression residual); y: dependent variable; $x_1,x_2,\ldots,x_i,\ldots,x_n$ are the independent variables; β_0: estimated ordinate at the origin; $\beta_0,\beta_1,\ldots,\beta_i,\ldots,\beta_n$ are the model coefficients.

2.2.2. ANN model

ANNs provide an alternative to mathematical modeling and they can be classified as nonparametric nonlinear models [45,46].

The neuron is the fundamental cell of an ANN, which can be considered as an elementary parallel operating processor. It is a computation unit which receives a number of inputs (x_i) directly from the environment or upstream neurons (Fig. 1). When information comes from a neuron, a weight is given to this latter which represents the ability of the neuron upstream to excite or inhibit the neuron downstream through its unique output passing by the activation function [47,48].

The neuron's output is calculated using Eq. (2):

$$S_J = f\left(\sum_{i=1}^{N} w_{ij} X_i + b_j\right) \tag{2}$$

where w_{ij}, is a synaptic weight, b_j is the bias input and X_i the ith input. f is the activation function which can usually be sigmoid or hyperbolic tangent [24].

Neurons can be connected together in a way to form a multilayer ANN. This latter is composed of an input layer, an output layer, and one or more hidden layers. All the neurons of a layer are connected to all neurons of the following layer through synaptic weights [49,50]. In Fig. 2, an example of an ANN with one hidden layer is given where each neuron (Fig. 1) is represented by a circle.

The goal is introduce to the input and output data of the ANN and make it learns the relationship between them by a process called learning using specific algorithms like backpropagation or Levenberg–Marquardt [51]. The goal is to minimize the error between the model output and the desired output by adjustment of synaptic weights.

To obtain the optimal structure of the neural network, we implement a strategy based on the design and optimization of the architecture of the neural network.

The development of the neural network model entails the following stages [24,52]:

(1) Collecting the experimental data.
(2) Define the input variables and the corresponding output variables.
(3) Pre-treatment and analysis of the data.
(4) Scaling and splitting of data for the phases of learning (with or without test) and validation.
(5) Selection of a neural network model.
The selection of the ANN model is affected by four major factors:
5-1: Network type (recurrent networks, feed-forward backpropagation, wavelet neural network, radial basis functions, etc.). In our work, we use the feed-forward back propagation neural network.
5-2: Network structure (number of hidden layers, number of neurons per hidden layer).
5-3: Activation functions.
5-4: learning algorithm. In our work, we use Levenberg–Marquardt learning (LM) algorithm.

First, transformation must be done in order to modify the distribution of input variables so that they can better match outputs. Before learning and validation, the inputs, and targets are scaled using a normalized equation (Eq. (3)) such that the data always fall within the interval [–1, 1] [53].

$$x_N = \left(y_{\max} - y_{\min}\right)\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) + y_{\min} \tag{3}$$

where x_N is the data value after normalization, $x_{\max}$ and $x_{\min}$ denote the maximum and the minimum of the data respectively; $y_{\max}$ and $y_{\min}$ are taken as 1 and –1; x denotes the data in question.

Data modeling was carried out by an ANN, 70% of the dataset, chosen randomly among the totality of the samples, were used for the learning phase. The remaining 30%, which did not participate in model learning, were divided into two parts (15% for the test and 15% for the validation) to examine the validity and performance of the prediction of these models [54].

2.3. Statistical evaluation criteria

The correlation coefficient (R), the adjusted coefficient R^2_{adj}, the root mean square error (RMSE), the mean square error (MSE), and the mean absolute error (MAE) were used out to estimate the performance of the model.

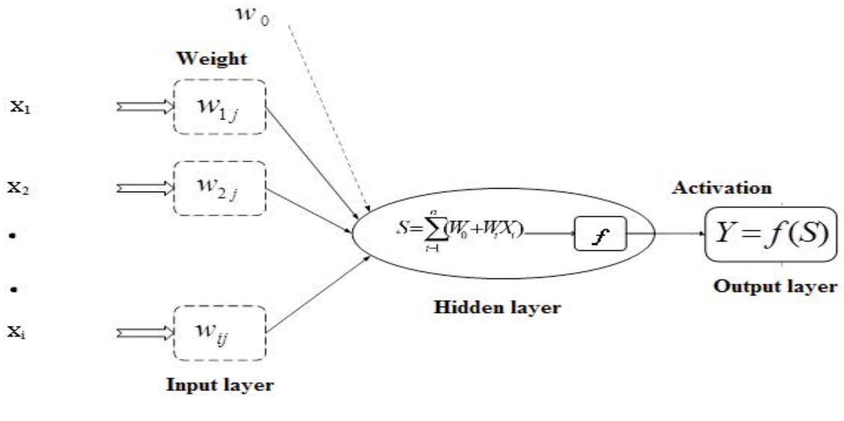


Fig. 1. Artificial neuron model.

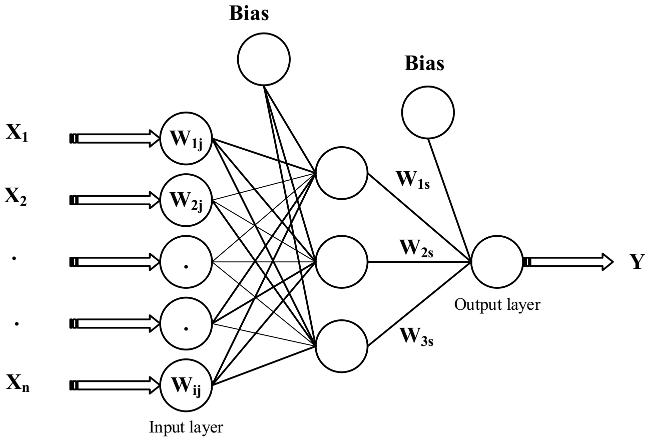


Fig. 2. Multilayer neural network.

The corresponding values are calculated using the equations [55,56]:

$$R = \frac{\sum_{i=1}^{N}\left(y_{\text{exp}} - \bar{y}_{\text{exp}}\right)\left(y_{\text{pred}} - \bar{y}_{\text{pred}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{\text{exp}} - \bar{y}_{\text{exp}}\right)^2 \sum_{i=1}^{N}\left(y_{\text{pred}} - \bar{y}_{\text{pred}}\right)^2}} \quad (4)$$

$$R^2_{\text{adj}} = 1 - \frac{\left(1 - R^2\right)\left(N - 1\right)}{N - K - 1} \quad (5)$$

$$\text{RMSE} = \sqrt{\left(\frac{1}{N}\right)\left(\sum_{i=1}^{N}\left[\left(y_{\text{exp}} - y_{\text{pred}}\right)\right]^2\right)} \quad (6)$$

$$\text{MSE} = \left(\frac{1}{N}\right)\left(\sum_{i=1}^{N}\left(y_{\text{exp}} - y_{\text{pred}}\right)^2\right) \quad (7)$$

$$\text{MAE} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left|y_{\text{exp}} - y_{\text{pred}}\right| \quad (8)$$

where N is the number of data; K is the number of variables (inputs); y_{ex} and y_{pr} are the experimental and the predicted values respectively; $\bar{y}_{\text{ex}}$ and $\bar{y}_{\text{pr}}$ are, respectively, the average values of the experimental and the predicted values [57–59].

The significance level value p and the F-ratio value which provide a measure of the statistical significance of the regression model were also determined. A high value of F with a minimum value of p means that the equation is significant [43].

3. Results and discussion

ANN and MLR are the two approaches, which were used for the prediction of the soluble sulfate concentration from the physicochemical parameters of drinking water. The two methods were performed and evaluated and then compared.

3.1. Multiple linear regression

A statistical analysis by the MLR method was performed using the "XLSTAT 2016" software on the database set. This method makes it possible to find the mathematical polynomial relation (Eq. (9)) between the soluble sulfate content and the independent variables, which corresponded to the physico-chemical parameters from 84 experimental samples. First, the equation takes into account all 18 variables (Table 1), even those that do not seem to have a significant impact on the dependent variable. The relation obtained was therefore evaluated in order to keep only the independent variables which were characterized by a high probability value (Table 1).

$$y = \beta_0 + \sum_{i=1}^{18}\beta_i x_i \quad (9)$$

where β_i and x_i are given in Table 1.

Then, only six independent variables which have a high power of explanation for the dependent variable were taken (P_r < 0.05); thus the relation can be reduced to Eq. (10):

$$y = \beta_0 + \sum_{i=6}^{18}\beta_i \times x_i \quad \text{for } i = 6,7,8,10,12 \text{ and } 18 \text{ only} \quad (10)$$

Indeed, the coefficients of each factor in the model make it possible to assess the impact of each factor on the response [60]. From Eq. (10), it is evident that changes in magnesium and sodium in drinking water increase the dose of sulfate. On the other hand, changes in chlorides, TAC, nitrogen dioxide, and bicarbonate dominate the sulfate dose.

The value of the coefficient of determination decreased slightly, but the equation became more simple after the elimination of low-explanatory variables from the dependent variable. The result of the learning phase shows an acceptable efficiency, via their correlation coefficient (Fig. 3) (R = 0.93795), but poor in terms of mean square error (RMSE = 84.0254 mg/L) and MAE (MAE = 48.6575 mg/L) (Table 2). The result of the learning phase was tested and validated by two databases, one for the test and the other for the validation. The results showed an acceptable correlation coefficient for the test phase (Fig. 3) (R = 0.92066) and a high correlation coefficient for the validation phase (Fig. 3) (R = 0.97100). However, in terms of RMSE (RMSE = 117.933 and 70.5674 mg/L) and absolute error mean (MAE = 61.8972 and 46.6106 mg/L), they always remained bad for the test and the validation phase, respectively (Table 2). The three steps were combined for the final evaluation of this model which gave an acceptable correlation coefficient (R = 0.941) (Fig. 3), but poor in terms of RMSE (RMSE = 88.3068 mg/L) and MAE (MAE = 42.3274 mg/L) (Table 2). In view of the obtained results, we can consider that the correlation of the model was a little positive via the acceptable correlation coefficient; while the statistical evaluation criteria remained high. The probability (P_r < 0.0001) was strictly less than 0.5% which confirms that the model was significant [11,43].

Results of MLR performances in terms of all errors and in terms of the agreement vector values (R, slope: α and y intercept: β) are given in Table 2.

Table 1
Values of β_i, x_i, and characterization of independent variables of the RLM

i	x_i	β_i	Standard error	T ratio	Prob. > |t|
0	Constant	245.4060	435.0382	0.56	0.5746
1	Conductivity (Cond)	–0.0576	0.0676	–0.85	0.3976
2	Turbidity (NTU)	4.3994	4.8354	0.91	0.3663
3	Potential hydrogen (PH)	–28.6495	55.9304	–0.51	0.6102
4	Hardness (TH)	0.0700	2.4603	0.03	0.9774
5	Calcium (Ca)	1.7242	0.8677	1.99	0.0511
6	Magnesium (Mg)	4.1885	1.2180	3.44	0.0010
7	Chlorides (Cl)	–1.1275	0.1141	–9.88	<0.0001
8	Total alkalimetric titer (TAC)	–0.7729	0.2281	–3.39	0.0012
9	Organic materials (MO)	0.4933	0.4757	1.04	0.3035
10	Nitrogen dioxide (NO_2)	–3,722.542	1,532.262	–2.43	0.0179
11	Nitrates (NO_3)	0.2332	2.4743	0.09	0.9252
12	Sodium (Na)	1.8031	0.2533	7.12	<0.0001
13	Potassium (K)	–3.3927	4.1309	–0.82	0.4145
14	Manganese (Mn)	–5,154.41	19,918.46	–0.26	0.7966
15	Iron (Fe)	135,558.938	18,425.29	0.74	0.4644
16	Aluminum (Al)	213.5702	447.0937	0.48	0.6345
17	Dry residues (Rs)	0.0828	0.0713	1.16	0.2500
18	Bicarbonate (HCO)	–0.6784	0.1533	–4.42	<0.0001

Table 2
Performances of the prediction using MLR

	RMSE (mg/L)	MAE (mg/L)	R	α	β
All	88.3068	42.3274	0.941	0.8859	46.480
Training	84.0254	48.6575	0.93795	0.9211	27.6758
Test	117.9331	61.8972	0.92066	0.7820	110.3162
Validation	70.5674	46.6106	0.97100	0.8464	77.2335

3.2. ANN Modeling

Preliminary tests showed that to improve the performance of a model established by ANN, it is necessary to modify the architecture of the network, by changing mainly the number of hidden layers, the number of hidden neurons, and/or the number of learning cycles (number of iterations). For this, we successively changed the number of hidden neurons (from 3 to 15). The results of these tests are shown in Table 3.

Table 3 presents the best architectures found. It shows the correlation coefficients and the error for each learning, test, and validation according to the number of neurons in the hidden layer and the network topology. It also indicates the activation functions for the hidden layer and the output layer. Architecture 3 (Table 3) appears to be the most relevant ANN model to predict the soluble sulfate content.

The network was driven until reaching over-learning; this phenomenon was met after 1,000 iterations. The over-learning phase was not yet reached, so it was interesting to continue learning until reaching this phase for the test in order to decrease the gradient further and thus improve the precision of the ANN. Using the third architecture of Table 3, the three curves related to the evolution of the MSE corresponding to the three learning phases were obtained (train, validation, and test) where they converge correctly toward a satisfactory minimum (Fig. 4).

Once the architecture of the ANN is synthesized, it is tested and the predicted values (estimated values) are compared to the experimental values in the three steps of learning phases; the corresponding results are given in Fig. 5.

The relation between the observed values and the estimated values shows the performance and efficiency of the developed ANN model. The relevance of the model was confirmed by the obtained coefficient of determination $R = 0.9996$ for the total data set.

3.3. Prediction performance

The result of the learning phase shows excellent efficiency, via their very high correlation coefficient (Fig. 6) (R = 0.99963), also very low values in terms of the MSE (RMSE = 6.5382 mg/L and the MAE (MAE = 4.2437 mg/L) (Table 4). The result of the learning phase was tested and

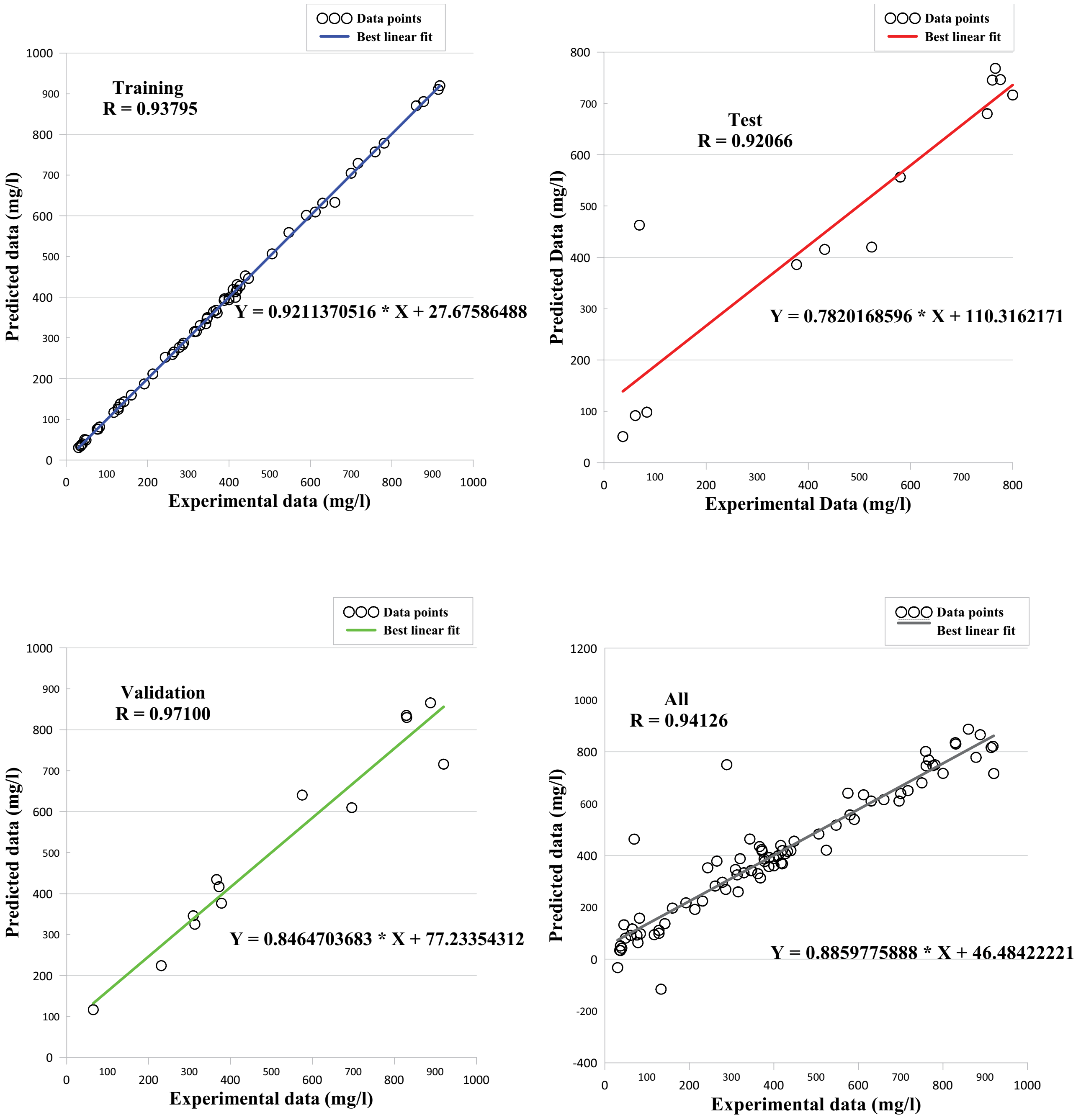


Fig. 3. Comparison between experimental and calculated values obtained by multiple linear regression.

Table 3
Performances of the different tested ANN architectures

ANN		Activation function		Coefficients of correlation				MSE		
Nbr	Neurons/ Layer	Hidden layer	Output layer	Learning	Test	Validation	Total	Learning	Test	Validation
1	[18–9–1]	tansig	purelin	0.99809	0.99964	0.99965	0.99863	2.34 × 10^{-4}	6.12 × 10^{-5}	7.15 × 10^{-5}
2	[18–13–1]	logsig	purelin	0.99961	0.99618	0.99723	0.99873	5.47 × 10^{-5}	5.27 × 10^{-4}	3.32 × 10^{-4}
3	[18–8–1]	tansig	tansig	0.99953	0.99978	0.99992	0.9996	7.36 × 10^{-5}	2.55 × 10^{-5}	9.79 × 10^{-6}

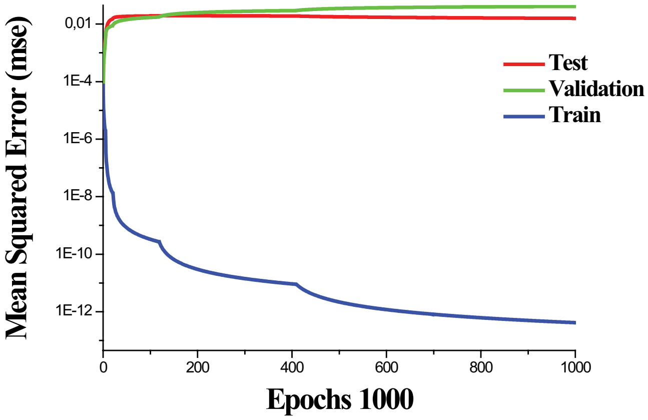


Fig. 4. Evolution of the MSE corresponding to the three learning phases.

validated by two databases, one for the test and the other for the validation; the result showed a very high correlation coefficient (*R* = 0.9998, 0.9999) for both phases (test and validation), respectively (Fig. 6), also very low values in terms of RMSE (RMSE = 5.4650 and 3.1826 mg/L) and MAE (MAE = 3.5362 and 1.9667 mg/L) for the test and the validation phases respectively (Table 4). The three steps were brought together for the final evaluation of this model which gave a very high correlation coefficient (*R* = 0.99973) (Fig. 6), and also led to very low values in terms of mean squared error (RMSE = 5.9755 mg/L) and MAE (MAE = 3.1767 mg/L) (Table 4). In view of the obtained results, we can consider that the correlation of the model was very positive, showing the efficiency of the model.

3.4. Interpolation performance

In order to test the precision of the ANN model previously developed and optimized, an interpolation was performed. For this purpose, a database was constructed in 2019 containing a set of data points located at the experimental points. The results showing the regression curve between the predicted and experimental values and the performance of the interpolation in terms of error and correlation coefficient are shown in Fig. 7.

These results show a good correlation between the predicted ANN and the experimental values with a very high correlation coefficient (Fig. 7) (*R* = 0.99918) and with a minimum squared error (RMSE = 9.3595 mg/L) and absolute error (MAE = 6.7598 mg/L) (Table 5). This result confirms again model efficiency using new data in the prediction range.

Results of interpolation performances in terms of all errors and in terms of the agreement vector values [*R* (correlation coefficient), α = (slope), and β (*y*-intercept)] are summarized in Table 5.

3.5. Extrapolation performance

Prediction was also performed to test the accuracy of the ANN developed. The idea was to test experimental data of the soluble sulfate content, which were not used during the training of our network; this database was built in 2019. The results in Fig. 8 show a very high coefficient (*R* = 0.9989), as well as very low values in terms of mean squared error (RMSE = 12.2587 mg/L) and MAE (MAE = 10.9054 mg/L) (Table 6). Extrapolation of a database was also achieved to check the accuracy of our optimized ANN model. The idea was testing experimental data sets never exploited during the learning and the test phases. The experimental data set of this system exploited in our principal database was obtained by analysis of water samples collected during several campaigns carried out in the Médéa region. Results of extrapolation performances in terms of all error and in terms of the agreement vector values [*R* (correlation coefficient), α = (slope), and β (*y*-intercept)] are summarized in Table 6. The quality of fit of the extrapolation data set is depicted in Fig. 8. The results showed a good agreement vector (*R*, α, and β) between experimental data and the ANN predicted results with accepted RMSE. It can be observed that for our extrapolation case, the results showed a good predictive ability of the ANN in both developed and optimized model. This shows a good convergence between the experimental output and the output predicted by the ANN.

3.6. Comparison of MLR and ANN for predicting soluble sulfate concentrations

In order to evaluate the two developed predictors, all data for MLR and ANN were compared (Table 7). The results depicted in Table 7 show the comparison of correlation coefficients and statistical indicators obtained by the two models. The correlation coefficient calculated by the ANN was significantly higher with (*R* = 0.99973) while that given by the MLR was less low (*R* = 0.941). In addition, the RMSE and MAE given by the ANN model (RMSE = 88.3068 mg/L and MAE = 42.3274 mg/L) were very low compared to that of the MLR model (RMSE = 5.9755 mg/L), (MAE = 3.1767 mg/L).

Fig. 9 confirms again the superiority of the ANN model if compared to the MLR model with a very good superposition of the curves measured experimentally and those estimated by the ANN model.

Table 4
Performances of the prediction using ANN

	RMSE (mg/L)	MAE (mg/L)	*R*	α	β
All	5.9755	3.1767	0.99973	1.0005	–0.1145
Training	6.5382	4.2437	0.99963	1.0018	–0.3500
Test	5.4650	3.5362	0.9998	0.9969	1.4545
Validation	3.1826	1.9667	0.9999	1.0024	–2.1524

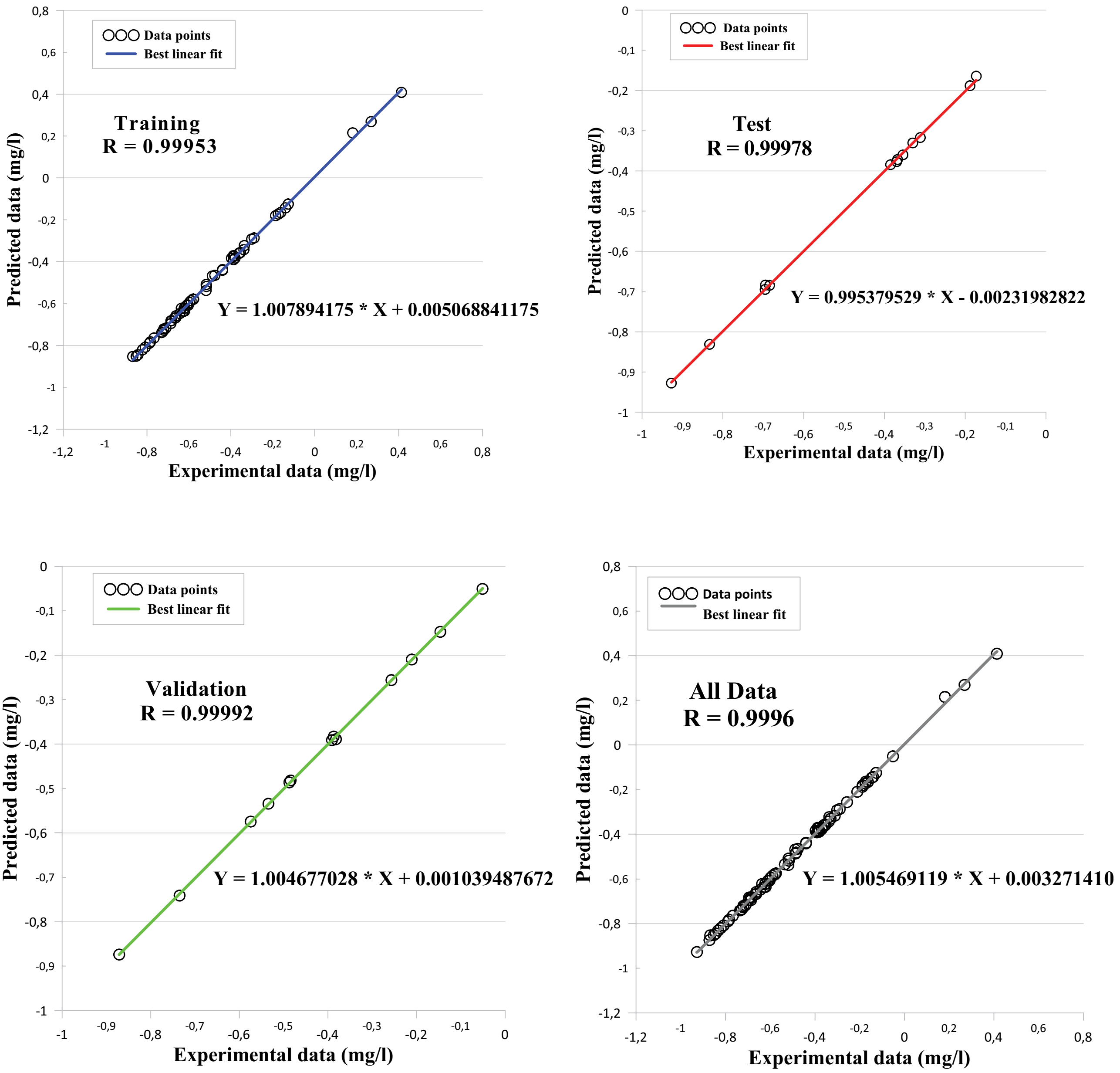


Fig. 5. Comparison between normalized experimental and calculated values for the ANN model.

The results obtained show a very good agreement for the ANN model, explained by a high correlation coefficient and small statistical indicators errors for the learning phase, the test phase, and the validation phase.

3.7. Residues study

The error made by the models established for each method on an individual of the model's construction sample is called residual [61]. Thus, the study of the relationship between the estimated levels of the soluble sulfate content by the mathematical models and their residuals ($y_{\text{exp}} - y_{\text{pred}}$) makes it possible to ensure the performance of the model and to verify empirically inter them, the validity of the hypotheses of the models.

The analytical methods to analyze the residues are mainly graphical analysis methods. Fig. 10 shows the residuals related to the model established by ANNs and those related to the model established by MLR based on the estimated values.

This figure shows that the residues obtained by the neural network method were less dispersed (close to zero) if compared to those obtained from the MLR.

In general, the results obtained were very satisfactory and justify the use of the neural network approach in the prediction of the soluble sulfate content levels of the Médéa region. It is consistent with the results of some recent studies that showed that MLR models perform poorly compared to those established by ANN models [48,54,62].

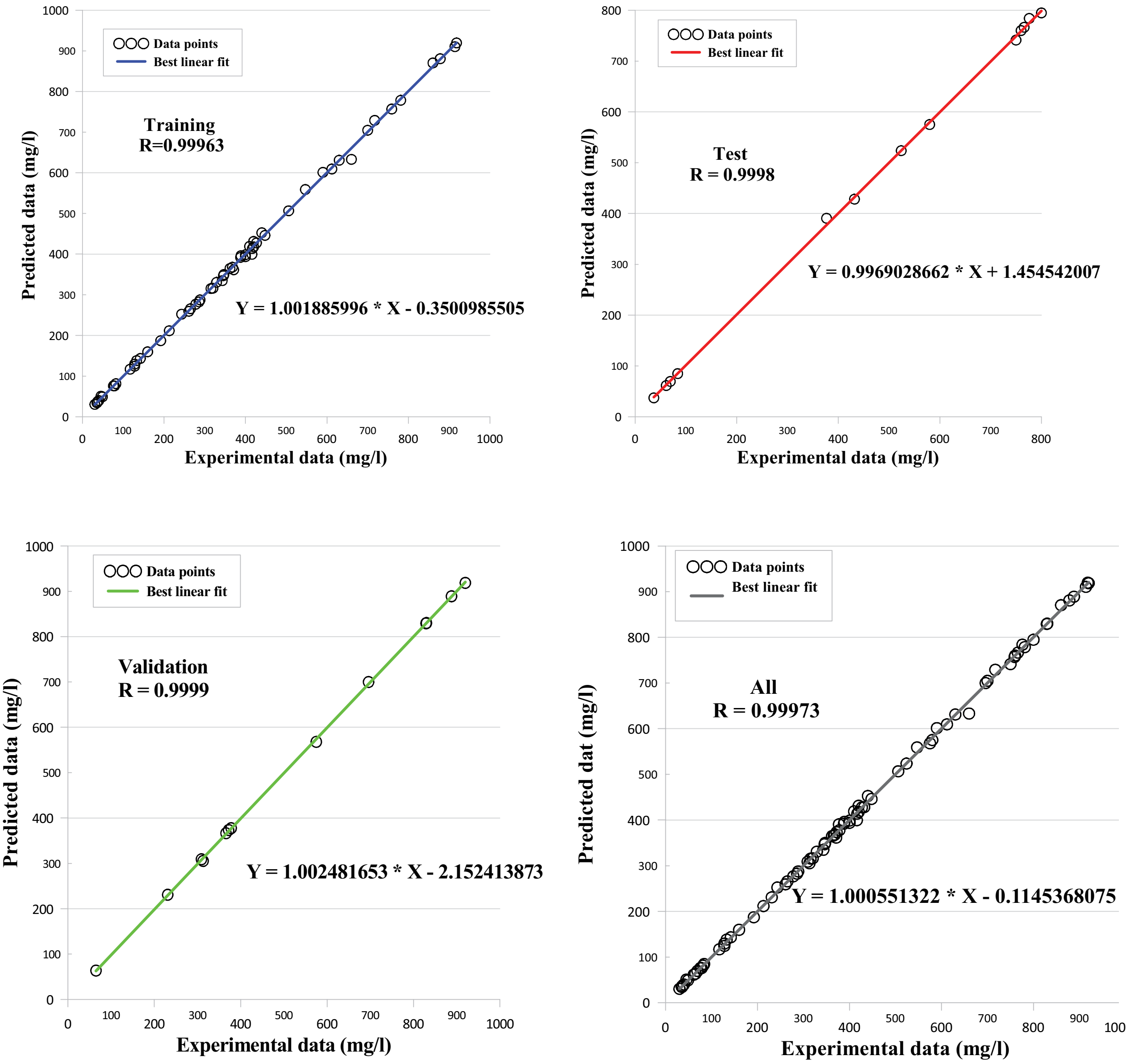


Fig. 6. Comparison between experimental and calculated values obtained by the ANN model to assess prediction performances.

Table 5
Performances of the interpolation

RMSE (mg/L)	MAE (mg/L)	*R*	α	β
9.3595	6.7598	0.99918	1.001	–7.027

Table 6
Performances of the extrapolation

RMSE (mg/L)	MAE (mg/L)	*R*	α	β
12.2587	10.9054	0.9989	1.0022	–4.6989

4. Conclusion

This document highlighted the importance of drinking water quality modeling based on the soluble sulfate content in the drinking water parameter. In this study, in the first stage, samples were collected at different points in the Medea region, Algeria. This was followed by a second stage, which was devoted to the analysis of these samples in the laboratory in order to have the physico-chemical parameters (sulfate, conductivity, turbidity, potential hydrogen, hardness, calcium, magnesium, chlorides, TAC, material organic, nitrogen dioxide, nitrates, sodium, bicarbonate, potassium, heavy metals (Mn^{2+}, Fe^{3+}, and Al^{+}), and dry residues). Then the database was constructed; statistical analysis was applied on physico-chemical parameters for the prediction of soluble sulfate content in drinking water in

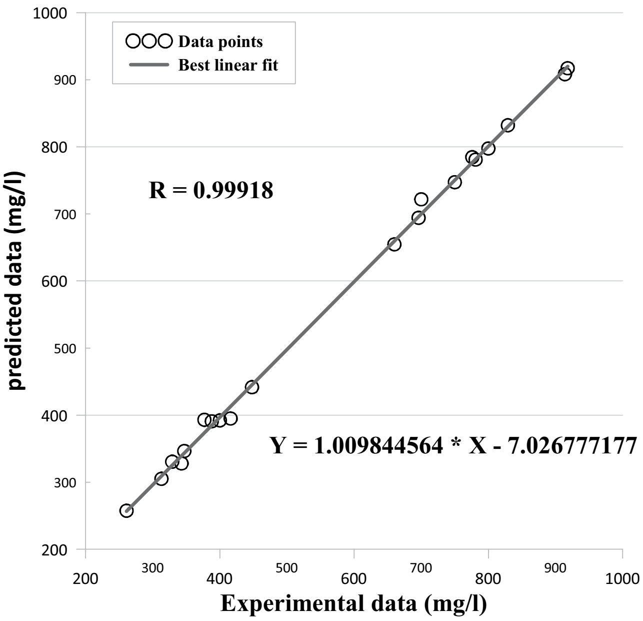


Fig. 7. Comparison between experimental and calculated values obtained by the ANN model to assess interpolation performances.

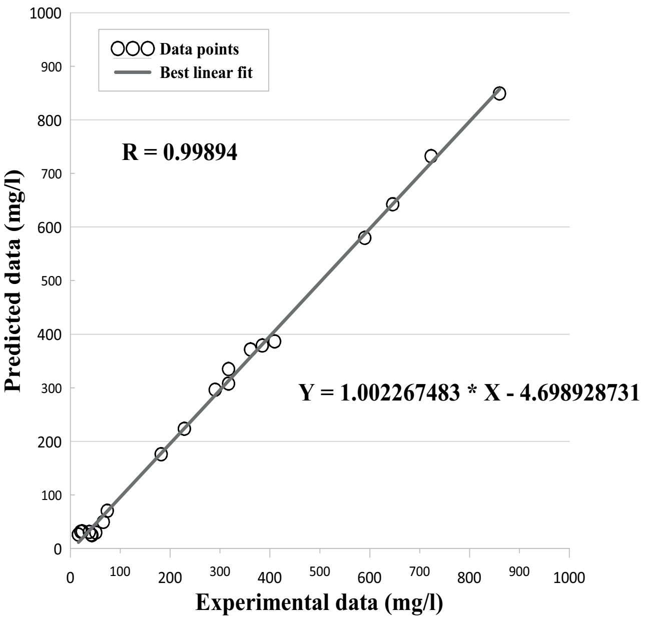
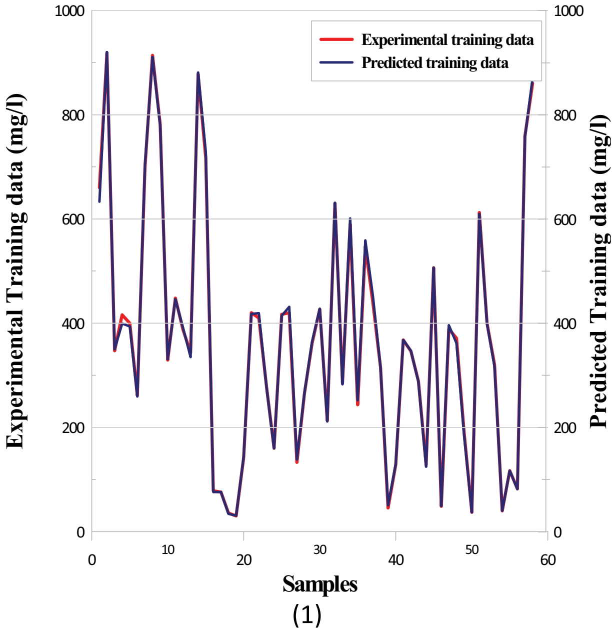


Fig. 8. Obtained by the ANN model to assess extrapolation performances.

Table 7
Performance comparison between MLR and ANN prediction models

Model	RMSE (mg/L)	MAE (mg/L)	R	α	β	R^2_{adj}
ANN	5.9755	3.1767	0.99973	1.0005	–0.1145	0.9996
MLR	88.3068	42.3274	0.941	0.8859	46.480	0.924



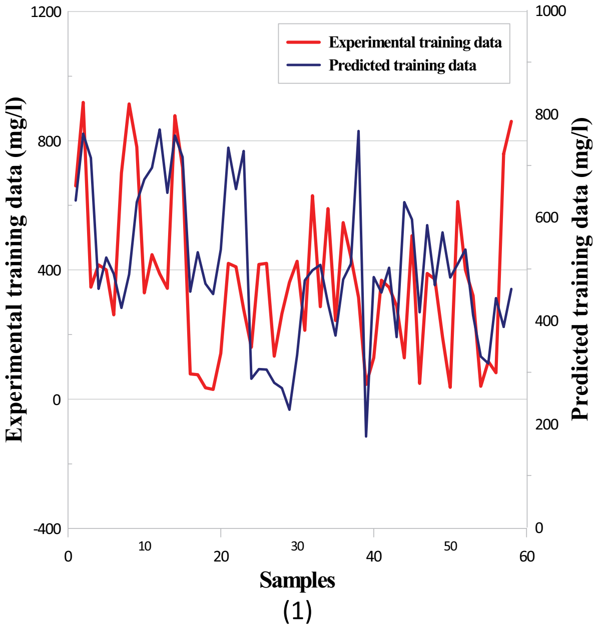


Fig. 9. Continued

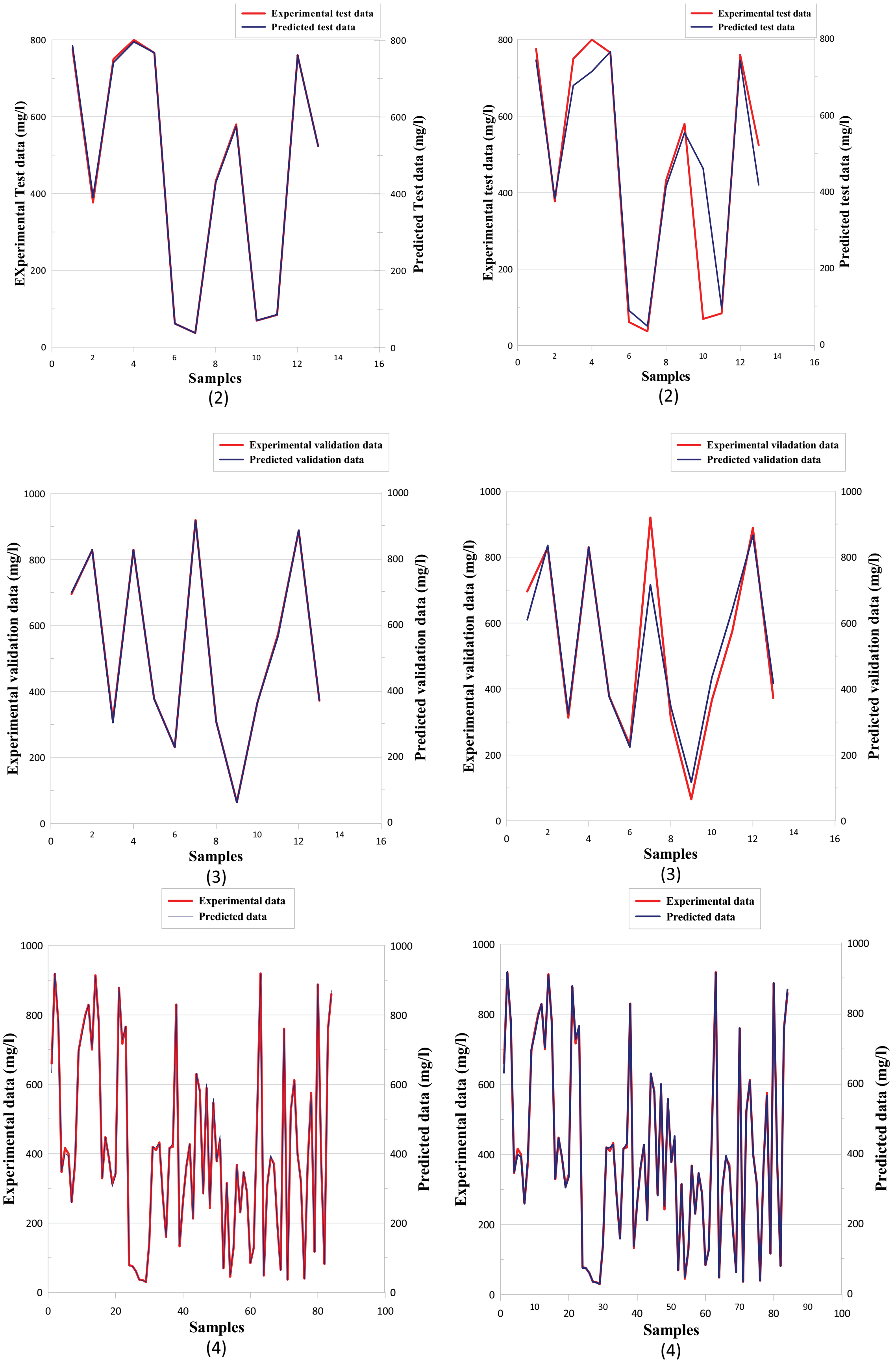


Fig. 9. Relationship between experimental data and the predicted data of samples using multiple linear regression and artificial neural network modeling: (1) training, (2) test, (3) validation, and (4) all data.

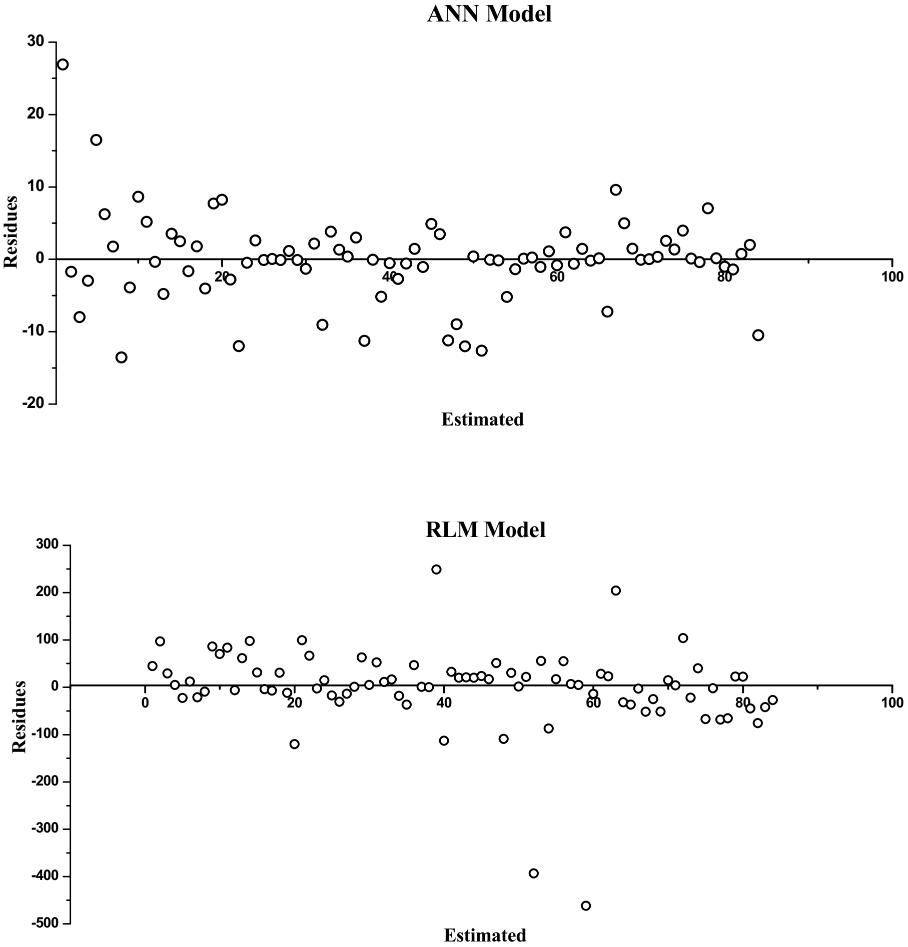


Fig. 10. Residues relating to the models established by multiple linear regression and artificial neural network depending on estimated values.

the Medea region using two modeling methods: MLR and neural networks for prediction of the soluble sulfate content in drinking water in order to compare between them. In addition, the RMSE and the MAE were used to evaluate the effectiveness of these two models.

The results showed that the predictive model based on ANNs of configuration [18–8–1] with hyperbolic tangent transfer functions in the hidden layer and in the output layer and the Levenberg-type learning algorithm Marquardt were better than those established by MLR. Indeed, the correlation coefficient of the ANN model was very high (R = 0. 99973) compared to the MLR model. Also, the RMSE, the RMSE, and the MAE of the ANN model were less than those established by MLR, 5.9755 and 88.3068 mg/L for RMSE, 35.7067 and 7,798.0935 mg/L for MSE, and 3.1767 and 42.3274 mg/L for MAE, respectively.

ANN model was re-tested considering two databases for interpolation and extrapolation to evaluate its efficiency. The results of the interpolation and extrapolation showed high efficiency, owing to their very high correlation coefficient and low values of the RMSE, the MSE, and the MAE. Consequently, the effectiveness of our model is confirmed. The results for interpolation and extrapolation were as follows: correlation coefficient R = 0.0.99918 and 0.9989, the RMSE (RMSE = 9.3595 and 12.2587 mg/L), the MSE (MSE = 87.6019 and 150.2771 mg/L), and the MAE (MAE = 6.7598 and 10.9054 mg/L), respectively.

This performance seems to be due to the fact that the soluble sulfate content was linked to the physico-chemical characteristics of the environment with non-linear relationships. In addition, the residue graphs showed the power of neural networks in the modeling of the data.

References

[1] X. Ding, Q. Zhu, A. Zhai, L. Liu, Water quality safety prediction model for drinking water source areas in Three Gorges Reservoir and its application, Ecol. Indic., 101 (2019) 734–741.

[2] K.W. Abdelmalik, Role of statistical remote sensing for Inland water quality parameters prediction, Egypt. J. Remote Sens. Space Sci., 21 (2018) 193–200.

[3] D.J. Booker, R.A. Woods, Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments, J. Hydrol., 508 (2014) 227–239.

[4] U. Seeboonruang, A statistical assessment of the impact of land uses on surface water quality indexes, J. Environ. Manage., 101 (2012) 134–142.

[5] M. Gevrey, L. Comte, D. de Zwart, E. de Deckere, S. Lek, Modeling the chemical and toxic water status of the Scheldt basin (Belgium), using aquatic invertebrate assemblages and

an advanced modeling method, Environ. Pollut., 158 (2010) 3209–3218.

[6] X. Xin, K. Li, B. Finlayson, W. Yin, Evaluation, prediction, and protection of water quality in Danjiangkou Reservoir, China, Water Sci. Eng., 8 (2015) 30–39.

[7] H. Runtti, S. Tuomikoski, T. Kangas, T. Kuokkanen, J. Rämö, U. Lassi, Sulphate removal from water by carbon residue from biomass gasification: effect of chemical modification methods on sulphate removal efficiency, Bioresources, 11 (2016) 3136–3152.

[8] C. Koschmann, A.-A. Calinescu, F.J. Nunez, A. Mackay, J. Fazal-Salom, D. Thomas, F. Mendez, N. Kamran, M. Dzaman, L. Mulpuri, ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma, Sci. Transl. Med., 8 (2016) 328ra28, doi: 10.1126/scitranslmed.aac8228.

[9] D. Guimarães, V.A. Leão, Batch and fixed-bed assessment of sulphate removal by the weak base ion exchange resin Amberlyst A21, J. Hazard. Mater., 280 (2014) 209–215.

[10] W. Chen, R. Zheng, P.D. Baade, S. Zhang, H. Zeng, F. Bray, A. Jemal, X.Q. Yu, J. He, Cancer statistics in China, 2015, Cancer J. Clin., 66 (2016) 115–132.

[11] A. El Hmaidi, H. El Badaoui, A. Abdallaoui, B. El Moumni, Application des réseaux de neurones artificiels de type PMC pour la prédiction des teneurs en carbone organique dans les dépôts du quaternaire terminal de la mer d'Alboran, Eur. J. Sci. Res., 107 (2013) 400–413.

[12] T. Rajaee, S. Khani, M. Ravansalar, Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review, Chemom. Intell. Lab. Syst., 200 (2020) 1–25, doi: 10.1016/j.chemolab.2020.103978.

[13] W. Deng, G. Wang, X. Zhang, A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting, Chemom. Intell. Lab. Syst. 149 (2015) 39–49.

[14] A.H. Haghiabi, A.H. Nasrolahi, A. Parsaie, Water quality prediction using machine learning methods, Water Qual. Res. J., 53 (2018) 3–13.

[15] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, Z. Miao, Water quality prediction model combining sparse auto-encoder and LSTM network, IFAC-PapersOnLine, 51 (2018) 831–836.

[16] H. Lu, X. Ma, Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere, 249 (2020) 1–12, doi: 10.1016/j.chemosphere.2020.126169.

[17] H.J. Mayfield, E. Bertone, C. Smith, O. Sahin, Use of a structure aware discretisation algorithm for Bayesian networks applied to water quality predictions, Math. Comput. Simul., 175 (2020) 192–201.

[18] A.N. Ahmed, F.B. Othman, H.A. Afan, R.K. Ibrahim, C.M. Fai, M.S. Hossain, M. Ehteram, A. Elshafie, Machine learning methods for better water quality prediction, J. Hydrol., 578 (2019) 1–18, doi: 10.1016/j.jhydrol.2019.124084.

[19] D. Wu, H. Wang, R. Seidu, Smart data driven quality prediction for urban water source management, Future Gener. Comput. Syst., 107 (2020) 418–432.

[20] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res., 171 (2020) 1–10, doi: 10.1016/j.watres.2019.115454.

[21] R. Avila, B. Horn, E. Moriarty, R. Hodson, E. Moltchanova, Evaluating statistical model performance in water quality prediction, J. Environ. Manage., 206 (2018) 910–919.

[22] E. Farahani, M.R. Mosaddeghi, A.A. Mahboubi, A.R. Dexter, Prediction of soil hard-setting and physical quality using water retention data, Geoderma, 338 (2019) 343–354.

[23] M. Khadr, M. Elshemy, Data-driven modeling for water quality prediction case study: the drains system associated with Manzala Lake, Egypt, Ain Shams Eng. J., 8 (2017) 549–557.

[24] H. Ousmana, A.E. Hmaidi, M. Berrada, B. Damnati, I. Etabaai, A. Essahlaoui, Development of a neural network approach for predicting nitrate and sulfate concentration in three lakes: Ifrah, Iffer and Afourgagh, Middle Atlas Morocco, Moroccan J. Chem., 6 (2018) 245–255.

[25] L.-M.L. He, Z.-L. He, Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA, Water Res., 42 (2008) 2563–2573.

[26] W.-C. Liu, W.-B. Chen, Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models, Comput. Geosci., 45 (2012) 13–25.

[27] A. Clementking, C.J. Venkateswaran, Prediction of Water Quality Attributes Variations Using Back Propagation Neural Network (BPNN) Model, International Conference on Technology and Business Management (ICTBM-15), American University in the Emirates, 2015, pp. 128–138.

[28] M. Heydari, E. Olyaie, H. Mohebzadeh, Ö. Kisi, Development of a neural network technique for prediction of water quality parameters in the Delaware River, Pennsylvania, Middle-East J. Sci. Res., 13 (2013) 1367–1376.

[29] H. Banejad, E. Olyaie, Application of an artificial neural network model to rivers water quality indexes prediction—a case study, J. Am. Sci., 7 (2011) 60–65.

[30] Y.R. Ding, Y.J. Cai, P.D. Sun, B. Chen, The use of combined neural networks and genetic algorithms for prediction of river water quality, J. Appl. Res. Technol., 12 (2014) 493–499.

[31] N.S. Jaddi, S. Abdullah, A cooperative-competitive master-slave global-best harmony search for ANN optimization and water-quality prediction, Appl. Soft Comput., 51 (2017) 209–224.

[32] A. Beucher, R. Siemssen, S. Fröjdö, P. Österholm, A. Martinkauppi, P. Edén, Artificial neural network for mapping and characterization of acid sulfate soils: application to Sirppujoki River catchment, southwestern Finland, Geoderma, 247 (2015) 38–50.

[33] N. Noori, L. Kalin, S. Isik, Water quality prediction using SWAT-ANN coupled approach, J. Hydrol., 590 (2020) 1–10, doi: 10.1016/j.jhydrol.2020.125220.

[34] S.S. Panda, V. Garg, I. Chaubey, Artificial neural networks application in lake water quality estimation using satellite imagery, J. Environ. Inf., 4 (2004) 65–74.

[35] H.Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, J. Environ. Health Sci. Eng., 12 (2014) 1–8, doi: 10.1186/2052-336X-12-40.

[36] J.-P. Suen, J.W. Eheart, Evaluation of neural networks for modeling nitrate concentrations in rivers, J. Water Resour. Plann. Manage., 129 (2003) 505–510.

[37] K. Ostad-Ali-Askari, M. Shayannejad, H. Ghorbanizadeh-Kharazi, Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran, KSCE J. Civ. Eng., 21 (2017) 134–140.

[38] S. Azimi, M.A. Moghaddam, S.H. Monfared, Prediction of annual drinking water quality reduction based on groundwater resource index using the artificial neural network and fuzzy clustering, J. Contam. Hydrol., 220 (2019) 6–17.

[39] F. Qaderi, E. Babanezhad, Prediction of the groundwater remediation costs for drinking use based on quality of water resource, using artificial neural network, J. Cleaner Prod., 161 (2017) 840–849.

[40] M.J. Diamantopoulou, D.M. Papamichail, V.Z. Antonopoulos, The use of a neural network technique for the prediction of water quality parameters, Oper. Res., 5 (2005) 115–125.

[41] J. Rodier, B. Legube, N. Merlet, R. Brunet, L'analyse de L'eau, 9e éd., Eaux Naturelles, Eaux Résiduaires, Eau de Mer, Dunod, 2009. Available at: https://books.google.dz/books?id=qUEGsUBZkL0C

[42] D. Jamin, Recherche du Boson de Higgs du Modèle Standard Dans le Canal de Désintégration ZH > nu nu bb Sur le Collisionneur Tevatron dans L'expérience D0. Développement D'une Méthode D'étiquetage des Jets de Quark b Avec des Muons de Basses Impulsions Transverses, 2010. Available at: https://tel.archives-ouvertes.fr/tel-00557839.

[43] M. Naoual, A. Abdelaziz, E.H. Abdellah, Use of artificial neural networks type MLP for the prediction of phosphorus level from the physicochemical parameters of sediments, IOSR J. Comput. Eng., 18 (2016) 61–70.

[44] A. Schmitt, B. Le Blanc, M.-M. Corsini, C. Lafond, J. Bruzek, Les reseaux de neurones artificiels. Un outil de traitement de données prometteur pour l'anthropologie, Bull. Mém. Soc. D'Anthropol. Paris, 13 (2001) 1–2.

[45] S. Huo, Z. He, J. Su, B. Xi, C. Zhu, Using artificial neural network models for eutrophication prediction, Procedia Environ. Sci., 18 (2013) 310–316.

[46] N.D. Kaushika, R.K. Tomar, S.C. Kaushik, Artificial neural network model based on interrelationship of direct, diffuse and global solar radiations, Sol. Energy, 103 (2014) 327–342.

[47] K.D. Fausch, C.L. Hawkes, M.G. Parsons, Models That Predict Standing Crop of Stream Fish from Habitat Variables: 1950-85, Gen. Tech. Rep. PNW-GTR-213 Portland US Department of Agriculture, Forest Service, Pacific, Northwest Research Station, 1988, 52 p.

[48] H. El Badaoui, A. Abdallaoui, I. Manssouri, L. Lancelot, Elaboration de modèles mathématiques stochastiques pour la prédiction des teneurs en métaux lourds des eaux superficielles en utilisant les réseaux de neurones artificiels et la régression linéaire multiple, J. Hydrocarbon Mines Environ. Res., 3 (2012) 31–36.

[49] E.M. Brakni, Réseaux de Neurones Artificiels Appliqués à la Méthode Electromagnétique Transitoire InfiniTEM, Université du Québec en Abitibi-Témiscamingue, 2011. Available at: https://depositum.uqat.ca/id/eprint/32

[50] R.P. Lippmann, An introduction to computing with neural nets, IEEE ASSP Mag., 4 (1987) 4–22.

[51] N. Samani, M. Gohari-Moghadam, A.A. Safavi, A simple neural network model for the determination of aquifer parameters, J. Hydrol., 340 (2007) 1–11.

[52] M. Sediri, S. Hanini, H. Cherifi, M. Laidi, S.A. Turki, Dynamic adsorption modelling of P-nitrophenol in aqueous solution using artificial neural network, J. Mater. Environ. Sci., 8 (2017) 2282–2287.

[53] R. Yacef, A. Mellit, S. Belaid, Z. Şen, New combined models for estimating daily global solar radiation from measured air temperature in semi-arid climates: application in Ghardaïa, Algeria, Energy Convers. Manage., 79 (2014) 606–615.

[54] H. El Badaoui, A. Abdallaoui, L. Lancelot, Application des réseaux de neurones artificiels et des régressions linéaires multiples pour la prédiction des concentrations des métaux lourds dans les sédiments fluviaux marocains, Eur. J. Sci. Res., 107 (2013) 400–413.

[55] D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, 1980. Available at: https://books.google.dz/books?id=ALjuAAAAMAAJ

[56] S.H. Hong, M.W. Lee, D.S. Lee, J.M. Park, Monitoring of sequencing batch reactor for nitrogen and phosphorus removal using neural networks, Biochem. Eng. J., 35 (2007) 365–370.

[57] I. Manssouri, M. Manssouri, B. El Kihel, Fault detection by K-NN algorithm and MLP neural networks in a distillation column: comparative study, J. Inf. Intell. Knowl., 3 (2011) 201–215.

[58] I. Manssouri, A. El Hmaidi, T.E. Manssouri, B. El Moumni, Prediction levels of heavy metals (Zn, Cu and Mn) in current Holocene deposits of the eastern part of the Mediterranean Moroccan margin (Alboran Sea), IOSR J. Comput. Eng., 16 (2014) 117–123.

[59] O.R. Dolling, E.A. Varas, Artificial neural networks for streamflow prediction, J. Hydraul. Res., 40 (2002) 547–554.

[60] S. Lefnaoui, N. Moulai-Mostefa, Investigation and optimization of formulation factors of a hydrogel network based on kappa carrageenan–pregelatinized starch blend using an experimental design, Colloids Surf., A, 458 (2014) 117–125.

[61] C. Voyant, M. Muselli, C. Paoli, M.-L. Nivet, Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation, Energy, 36 (2011) 348–359.

[62] M. Bélanger, N. El-Jabi, D. Caissie, F. Ashkar, J. Ribi, Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple, J. Water Sci., 18 (2005) 403–421.