Assessment of water quality index in unmonitored river basin using multilayer perceptron neural networks and principal component analysis

Bachir Sakaa\textsuperscript{a,b,*}, Nabil Brahmia\textsuperscript{c}, Hicham Chaffai\textsuperscript{b}, Azzedine Hani\textsuperscript{b}

\textsuperscript{a}Centre de Recherche Scientifique et Technique sur les Régions Arides CRSTRA, BP 1682 RP, 07000 Biskra, Algérie, email: sakaabachir@yahoo.fr (B. Sakaa)
\textsuperscript{b}Laboratoire Ressource en Eau et Développement Durable, Faculté des Sciences de la Terre, Université Badji Mokhtar, BP 12, 23000 Annaba, Algérie, emails: hichamchaffai@yahoo.fr (H. Chaffai), haniazzedine@yahoo.fr (A. Hani)
\textsuperscript{c}Laboratoire des Réservoirs Souterrains Pétroliers, Gaziers et Aquifères, Ouargla, Algeria, email: nabilbra@yahoo.fr (N. Brahmia)

Received 26 November 2019; Accepted 12 May 2020

\textbf{Abstract}

This study describes the combination of neural networks and multivariate methods to develop a proper model for the forecasting of water quality index (WQI) in the Saf-Saf River using water quality parameters. The main objectives of this work were to determine the importance of different input variables and to assess the spatial and temporal water quality variation. MLP models were trained using three different algorithms and tested, these models were compared in terms of efficiency criteria and goodness-of-fit for WQI modeling. The results show that MLP\textsubscript{BFGS} model provide the best performance with small root mean square error value (RMSE = 0.007) and high coefficient of determination value ($R^2 = 0.811$) compared with the other types of MLP models. In the meantime, sensitivity analysis reveals that BOD\textsubscript{5} acts as the most contributor decreasing WQI. PCA/FA results show relatively spatial and seasonal changes in surface water quality, it generated three groups of sampling sites with similar characteristics. Group I (upstream sites), group II (midstream sites), and group III (downstream sites) correspond to a relatively low pollution, moderate pollution, and high pollution sites, respectively. Therefore, this approach can provide managers with the right tools to make decisions about the implementation of sustainable management practices.

\textbf{Keywords:} Water quality index; Multilayer perceptron; Principal component analysis; Factor analysis; Saf-Saf river basin

\section{1. Introduction}

Issues and challenges related to water quality deterioration have generated much debate and discussion on heightening awareness on water quality concerns, and the increasing demand to sustainably manage our water resources. In the last decades, a rapid industrial development without controlling discharges, the intensive use of fertilizers in agriculture and the over exploitation of water resources destroys river ecosystems and affect human health in different ways. This event produces a chemical modification of the water rendering it unusable for other purposes and hence aggravates scarcity of water resources [1].

Surface water resources is unfortunately exposed more and more to pollution, in the form of discharges of industrial or domestic effluents, and are gradually becoming unfit for any use, without prior treatment. The latter is often complicated and expensive [2,3]. Guaranteeing a good water supply is not enough anymore, it must also be avoided that after-use water, known as wastewater, contaminates groundwater, rivers, and lakes, thus rendering them unfit for consumption and industrial use. It is, therefore, becoming
increasingly necessary to contribute to a dual program of conserving and protecting water. For that, a better knowledge of the analytical level of the pollution of the rivers is essential [4].

Prevention of river pollution requires effective monitoring of physicochemical and biological parameters [5]. The water quality index (WQI) is a means of summarizing large amounts of water quality data into simple terms (e.g., good, fair, and poor) for reporting to policymakers and the public in a comprehensive, consistent manner [6]. The WQI is used to state the pollution status of hydro-systems, because it represents a single numeric score that describes the water quality condition at a particular location in a specific time [7,8]. There are different approaches of WQI were used by various countries and institutions around the world to assess the water quality status of their rivers like Argentina [9]; USA [10]; India [11]; Portugal [12]; Turkey [13]; and China [14,15]. In addition, water quality information becomes more easily and quickly interpretable than a list of numeric values.

Recently, several actors intervening at different levels in water resources management of the Saf-Saf river basin, (water quality specialists, other managers, legislators, or the general public), need to analyze and process this information in order to effectively fulfill their role. It becomes so imperative to simplify this perception of water quality so that the extension of water quality can serve a technical, social, and/or even political purpose [16,17]. The objectives of this study, therefore, are two-fold; the first objective is to develop WQI for assessing surface-water quality and defining water pollutants. The second objective is to establish a proper model based on artificial neural networks and multivariate statistical techniques; this model is aimed to assist planners and managers of water resources systems for solving surface water pollution problems. The ANN model provides a perfect knowledge and understanding about the relationship between water quality parameters and WQI, and defines the effective water quality parameter influencing the decreasing values of WQI through different sampling sites in Saf-Saf river basin.

Artificial neural networks (ANNs) have been successfully applied in a number of diverse fields including water resources [18,19]. In the context water quality prediction, ANNs may offer a promising alternative for water quality parameters [20–23]. There are many published works in the field of wastewater treatment plant performance using artificial intelligence methods such as neural networks [24,25]. Clair and Ehrman [26] used 10 y of data to examine the relationships between climate and geography on discharge and dissolved organic carbon (DOC) and dissolved organic nitrogen (DON) from 15 rivers in Canada’s Atlantic region. Karul et al. [27] used a three-layer Levenberg–Marquardt feedforward learning algorithm to model the eutrophication process in three water bodies of Turkey (Keban Dam Reservoir, Mogan, and Eymir Lakes). Zhao et al. [28] developed three-layer feed-forward neural networks with back propagation (BP) for predicting biochemical oxygen demand (BOD) in Yujiao Reservoir, China with a correlation coefficient of 0.85537 and average error of 2.56%.

Multivariate statistical techniques including principal component analysis (PCA) and factor analysis (FA) has been widely applied in environmental data reduction and interpretation of multiconstituent chemical, physical, and biological measurements [29,30]. Both PCA and FA are very powerful techniques whose main objective is to reduce the dimensions of a multivariate data set [2]. In addition, it allows to evaluate the relationship between variables, since they show the contribution of individual chemicals in several influence factors [31], also Helena et al. [32] using multivariate statistical techniques to characterize and evaluate groundwater quality, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality.

In our present study, we have applied artificial neural networks ANNs to forecasting WQI in the Saf-Saf river basin based on a cause–effect relationship. Here, we have investigated the possibility of building a relationship between water quality parameters (independent variables) with WQI (dependent variable). What’s more, the ANNs and decision-makers opinion are used in the characterizing and prioritizing of the most effective variable. The selected variables have been classified using the multivariate statistical techniques including principal components analysis and factor analysis.

2. Materials and methods

2.1. Study area and data description

Saf-Saf river basin is located on North-East of Algeria between parallels 6°40′–7°10′ East and 36°25′–36°53′ North, this basin covers a surface of 1,158 km², limited by the Gueblı River basin from the west, the Hajar Mountain from the south, Kebir West river basin from the east, and finally the Mediterranean Sea from the north (Fig. 1). The climate is sub-humid with average annual rainfall varies from 636 mm in South to 750 mm in North, and the average monthly temperatures (minimal and maximal) varied between 12°C and 36°C [16].

The components of water resources balance for the Saf-Saf river basin has been developed based on the estimates of all water inputs and outputs to the river basin. Table 1 shows that the present net water balance in the Saf-Saf river basin is negative (–6.28 hm³ y⁻¹) which indicates that there is a water deficit. The negative balance leads to decreasing the volume of freshwater in the river basin and the degradation of water quality [17].

A total of 35 samples of surface water were collected at various sampling sites along Saf-Saf river basin (Fig. 1). During April and September, 2015; all the water samples were sampled from a depth of 15 cm below the surface and preconditioned high density polyethylene bottles. They were conditioned by washing initially with five percent (5%) nitric acid, and then rinsing several times with distilled water. This was carried out to ensure that the sampling bottles were free from contaminants. Each of the surface water samples was analyzed for various physicochemical and biochemical parameters such as water temperature (WT, °C), the potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L⁻¹), turbidity (NTU), nitrate (NO₃, mg L⁻¹), phosphate (PO₄³⁻, mg L⁻¹), 5 d BOD (mg L⁻¹), chemical oxygen demand (COD, mg L⁻¹), and chloride (Cl, mg L⁻¹).
The WQI values are calculated using the software CCME calculator version 1.0 developed by Canadian Council of Ministers of the Environment [6]. The CCME WQI was originally developed as the Canadian Water Quality Index (CWQI). It comprises of three factors and is well-documented [6]:

- **Factor 1 (scope):** represents the percentage of variables that do not meet their objectives at least once during the time period under consideration (failed variables), relative to the total number of variables measured:

\[
F_1 = \left( \frac{\text{Number of failed variables}}{\text{Total number of variables}} \right) \times 100
\]  

- **Factor 2 (frequency):** represents the percentage of individual tests that do not meet objectives ("failed tests"): 

\[
F_2 = \left( \frac{\text{Number of failed tests}}{\text{Total number of tests}} \right) \times 100
\]  

- **Factor 3 (amplitude):** represents the amount by which failed test values do not meet their objectives. \( F_j \) is calculated in three steps.

The number of times by which an individual concentration is greater than (or less than, when the objective is a minimum) the objective is termed an "excursion" and is expressed as follows. When the test value must not exceed the objective:

\[
\text{Excursion}_j = \left( \frac{\text{Failed test value}_j}{\text{Objective}_j} \right) - 1
\]  

For the cases in which the test value must not fall below the objective:

\[
\text{Excursion}_j = \left( \frac{\text{Objective}_j}{\text{Failed test value}_j} \right) - 1
\]  

- The amount by which individual tests are out of compliance is calculated by summing the excursions of
individual tests from their objectives and dividing by the total number of tests (both those meeting objectives and those not meeting objectives). This variable referred to as the normalized sum of excursions, or nse, is calculated as:

\[
\text{nse} = \frac{\sum \text{Excursion}}{\text{Number of tests}}
\]

(5)

- \( F_3 \) is then calculated by an asymptotic function that scales the normalized sum of the excursions from objectives (nse) to yield a range between 0 and 100.

\[
F_3 = \left( \frac{\text{nse}}{0.01\text{nse} + 0.01} \right)
\]

(6)

Once the factors have been obtained, the index itself can be calculated by summing the three factors as if they were vectors. The sum of the squares of each factor is therefore equal to the square of the index. This approach treats the index as a three-dimensional space defined by each factor along one axis. With this model, the index changes in direct proportion to changes in all three factors [6].

\[
\text{WQI} = \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732}
\]

(7)

The divisor 1.732 normalizes the resultant values to a range between 0 and 100, where 0 represents the worst water quality and 100 represents the best water quality. Once the WQI value has been calculated, water quality is ranked by relating it to one of the following classes (Table 2).

### 2.2. Methodology

In this research, ANNs, decision-makers opinion and judgment, descriptive statistics and multivariate statistical techniques were used in the characterizing of WQI [33–35]. The analysis plan may be decomposed into four major steps which again are decomposed into many tasks (Fig. 2). The contents of the four steps are:

- **Step 1**: the first step aims to create a neural networks model, characterize, and prioritize the effective water quality parameters, and to establish a relationship between water quality parameters and WQI.
- **Step 2**: this step expresses the analysis of the questionnaire data to examine the decision-makers opinion and judgment of various stakeholders using descriptive statistics. The results of step 2 were compared with the results of the ANNs in step 1 to explore the understanding and knowledge of the local decision-makers about the health status of surface water in Saf-Saf river basin.
- **Step 3**: the purpose of this step is to transform the variables that were not normally distributed and to calculate the correlation matrix the variables selected from step 1.
- **Step 4**: Two methods of multivariate statistical techniques (PCA and FA) were used in step (4) for the selected water quality parameters, to classify them with the different sampling sites during wet and dry season.

<table>
<thead>
<tr>
<th>Class</th>
<th>Sampling sites</th>
<th>WQI value</th>
<th>Water status and observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>/</td>
<td>95–100</td>
<td>Excellent: water quality is protected with a virtual absence of threat or impairment; conditions very close to natural levels</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>80–94</td>
<td>Good: water quality is protected with only a minor degree of threat or impairment; conditions rarely depart from natural or desirable levels</td>
</tr>
<tr>
<td>III</td>
<td>MS1, MS3, MS5,</td>
<td>65–79</td>
<td>Fair: water quality is usually protected but occasionally threatened or impaired; conditions sometimes depart from natural or desirable levels</td>
</tr>
<tr>
<td></td>
<td>MS10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>MS1, MS2, MS3,</td>
<td>45–64</td>
<td>Marginal: water quality is frequently threatened or impaired; conditions often depart from natural or desirable levels</td>
</tr>
<tr>
<td></td>
<td>MS4, MS5, MS6,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS6, MS7, MS8,</td>
<td></td>
<td>Poor: water quality is almost always threatened or impaired; conditions usually depart from natural or desirable levels</td>
</tr>
<tr>
<td></td>
<td>MS9, MS10, MS11,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS11, MS12, MS13,</td>
<td>0–44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS14, MS15, MS16,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>MS17, MS18, MS19,</td>
<td>0–44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS20, MS21, MS22,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS23, MS24, MS25,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS26, MS27, MS28,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS29, MS30, MS31,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS32, MS33, MS34,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS35, MS36, MS37,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MS1: sampling site during wet season and MS1: sampling site during dry season
2.2.1. Artificial neural networks

ANNs are able to map input–output relationships for natural complex phenomena and were developed to model the brain's interconnected system of neurons so that computers could use to imitate the brain's ability to sort patterns and learn from trial and error, thus observing relationships in data [36]. The main differences between various types of ANNs involve network architecture, method for determining the weights and transfer function as a creator of output value [37]. Feed-forward neural networks with back propagation are successfully applied to environmental problems. Multilayer perceptron (MLP) is perhaps the most popular network architecture in use today, due originally to Rumelhart et al. [38] and discussed at length in most neural network textbooks [39].

In this study, three-layer feed-forward MLP neural networks with gradient descent (GD), Broyden–Fletcher–Goldfarb–Shanno (BFGS), and conjugate gradient(CG) back-propagation learning were developed for the relationship between water quality parameters and WQI. The variables representing the water quality parameters were considered as the possible input variables including water temperature (WT, °C), potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L\(^{-1}\)), turbidity (NTU), nitrate (NO\(_3\)\(^-\), mg L\(^{-1}\)), phosphate (PO\(_4\)\(^3-\), mg L\(^{-1}\)), 5 d BOD (mg L\(^{-1}\)), COD (mg L\(^{-1}\)), and chloride (Cl\(^-\), mg L\(^{-1}\)), while the target output variable was the WQI, which is the major means of assessing the levels of pollution of the Saf-Saf river. The MLP Neural network can be represented by the following compact form:

\[
\{\mathrm{WQI}\} = \text{ANN}_{\text{MLP}}\left\{\text{pH, WT, OS, TDS, Turbidity, NO}_3, \text{PO}_4^{3-}, \text{BOD}_5, \text{COD, Cl}^-\right\}
\]

A schematic diagram of MLP neural network is given in Fig. 3. It shows a typical feed forward MLP structure with signals flow from input nodes, forward through hidden nodes, eventually reaching the output node.

Each hidden node \((i)\) receives signals from every input node \((i)\) which carries standardized values \((X)\) of an input variable where various input variables have different measurement units and span different ranges. \(X_i\) is expressed as:

\[
\bar{X}_i = \frac{X_i - X_{\min}(i)}{X_{\max}(i) - X_{\min}(i)}
\]

Each signal comes via a connection that has a weight \((W_j)\). The net integral incoming signals to a receiving hidden node (Net) is the potential of the neuron, \(\bar{X}_i\) and the corresponding weights, \((W_j)\) plus a constant reflecting the node threshold value (TH):

\[
\text{Net}_j = \sum_{i=1}^{n} \bar{X}_i W_j + \text{TH}_j
\]
where \((O_j)\) passes as a signal to the output node \((k)\). The net entering signals of an output node (\(Ne_t\)):

\[
Ne_t = \sum_{i} O_j W_{ji} + TH_j
\]

The net incoming signals of an output node (\(Net_i\)) transformed using the sigmoid type function to a standardized or scaled output \((\bar{O}_i)\) that is:

\[
\bar{O}_i = f(Net_i) = \frac{1}{1 + e^{-Net_i}}
\]

Then, \((\bar{O}_j)\) is standardized to produce the target output:

\[
O_k = \bar{O}_j \left[ O_{\text{max}} - O_{\text{min}} \right] + O_{\text{min}}
\]

According Rumelhart et al. [38], the sigmoid function must be continuous, differentiable, and bounded from above and below in the range \([0,1]\). The calculated error between the observed value and the simulated value of the dependent variable is back propagated through the network and the weights are adjusted. Liu et al. [40] confirmed that the cyclic process of feed forward and error back propagation are repeated until the validation error is minimal.

The performance of each of the selected models (MLP<sub>DG</sub>, MLP<sub>CIG</sub>, and MLP<sub>BFGS</sub>) was determined using the criteria, such as the root mean square error (RMSE), the coefficient of determination \((R^2)\), and the accuracy factor \((A)\) computed from the measured and model predicted values of the dependent variables [41,42]. Values of the criteria parameters were calculated for all the two sets (training and test) as:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (WQI_i - \hat{WQI}_i)^2}{N}}
\]

\[
R^2 = \frac{\sum_{i=1}^{N} \left( WQI_i - WQI \right)^2}{\sum_{i=1}^{N} \left( WQI_i - \hat{WQI}_i \right)^2}
\]
Table 3
Basic statistics of water quality parameters during wet and dry season (N = 35)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Wet season</th>
<th>Dry season</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Minimum</td>
</tr>
<tr>
<td>pH</td>
<td>7.50</td>
<td>7.24</td>
</tr>
<tr>
<td>WT</td>
<td>20.27</td>
<td>15.50</td>
</tr>
<tr>
<td>OS</td>
<td>84.81</td>
<td>21.40</td>
</tr>
<tr>
<td>TDS</td>
<td>369.1</td>
<td>60.00</td>
</tr>
<tr>
<td>Turbidity</td>
<td>59.00</td>
<td>5.00</td>
</tr>
<tr>
<td>NO₃⁻</td>
<td>2.20</td>
<td>0.16</td>
</tr>
<tr>
<td>PO₄³⁻</td>
<td>0.94</td>
<td>0.003</td>
</tr>
<tr>
<td>BOD₅</td>
<td>14.58</td>
<td>1.80</td>
</tr>
<tr>
<td>COD</td>
<td>23.75</td>
<td>2.96</td>
</tr>
<tr>
<td>Cl⁻</td>
<td>95.21</td>
<td>25.00</td>
</tr>
<tr>
<td>WQI</td>
<td>54.54</td>
<td>29.00</td>
</tr>
</tbody>
</table>

where WQI is the observed output value; WQI is the simulated output value; WQI is the mean value of WQI values; N is the total number of data sets. The RMSE, a measure of the goodness-of-fit, best describes an average measure of the error in predicting the dependent variable. $R^2$ value is an indicator of how well the network fits the data and accounts for the variability with the variables specified in the network [43]. A value of $R^2$ above 90% refers to a very satisfactory model performance. The $A_f$ is a simple multiplicative factor showing the spread of simulation results. The larger value of $A_f$ the less accurate is the average estimation. A value of 1 indicates that there is a perfect agreement between all the predicted and the measured values. Finally, the goodness-of-fit of the selected models (MLP-DG, MLP-CG, and MLP-BFGS) were also checked through the analysis of the residuals.

2.2.2. Correlation matrix

Correlation matrix is a table evaluating the relationship between water quality variables. It calculates the direction and strength of the relationship between any two variables in the data set. A correlation coefficient near -1 or 1 means the strongest negative or positive relationship between two variables and its value closet to 0 means no linear relationship between them at a significant level of $p < 0.05$ [44]. The most commonly used measure of correlation is Pearson’s $r$, it is called the linear correlation coefficient because $r$ measures the linear association between two variables. Pearson’s $r$ assumes that the data follow bivariate normal distribution [46]. The correlation coefficient can be used to estimate the population Pearson correlation $r$ between $X$ and $Y$, it is written as:

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1)s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

(18)

2.2.3. Multivariate statistical techniques (PCA and FA)

Principal component analysis is a multivariate technique focused on a particular collection of variables. It is a powerful tool for pattern recognition that explains the variance of a large set of inter-correlated variables and transforms them into a smaller set of independent principal components [2,32]. These PCs provides information on the most meaningful parameters, which describe the whole data set through data reduction with a minimum loss of original information. Each principal component (PC) is a linear combination of the original variables and describes a different source of variation. PC is expressed as:

$$PC_i = w_1 x_1 + w_2 x_2 + \ldots + w_u x_u$$

(19)

where $x_i$ and $w_i$ are the original variable and the component weight, respectively. The principal component weights are used as measures of the correlation between the variables and the principal components. The special feature of PCA is the graphics that provide a visual aid for the classification of variables and cases.

Factor analysis (FA) is formulated to transform the original variables into new uncorrelated variables called factors, which are linear combinations of the original variables. In addition, during the computation of FA, the most researchers performed a varimax rotation (raw) of the principal components coming from the original standardized variables, in order to reduce the contribution of variables with minor significance. The Varimax rotation was done taking into account previous works using FA for the evaluation of temporal and spatial changes in water quality [31,32]. The FA can be expressed as:

$$Z_{ij} = a_1 x_{ij} + a_2 x_{ij} + \ldots + a_n x_{nj} + e_{ij}$$

(20)
where $Z$ is the measured variable, $f$ is the factor score and $e$ is the residual term accounting for errors or other source of variation.

### 3. Results and discussion

#### 3.1. Summary descriptive statistics of water quality parameters and WQI

In this research, the data sets that we analyzed from 35 sampling sites in the study area were processed. The selected parameters for the estimation of surface water quality characteristics were: water temperature (WT, °C), potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L⁻¹), turbidity (NTU), nitrate (NO₃, mg L⁻¹), phosphate (PO₄³⁻, mg L⁻¹), 5 d BOD₅ (mg L⁻¹), COD (mg L⁻¹), chloride (Cl⁻, mg L⁻¹), and water quality index (WQI, %). The summarized basis statistics of these parameters (minimum, maximum, standard deviation, and skewness) are presented in Table 3.

#### 3.2. Artificial neural networks

The main dataset is divided into two sub-datasets: training (used 80% of the total available sets for finding the appropriate weight for each input) and the test dataset (20% of the total available sets for the evaluation of actual model performance). The test datasets were extracted randomly. Different MLP models (MLP (CG 45), MLP (CG 39) and MLP (BFGS)) were created and tested in order to determine the optimum number of nodes in the hidden layer. According to Fletcher and Goss [45], the appropriate number of nodes in a hidden layer ranges from $(2n^{1/2} + m)$ to $(2n + 1)$, where $n$ is the number of input nodes and $m$ is the number of output nodes. Regarding the results obtained from the 100 MLP models created using three different algorithms, it can be concluded that the best optimal MLP model found is MLP (BFGS) with 14 hidden nodes and a minimal root mean square error RMSE of 0.009 in testing data sets compared with the other types of MLP networks (Table 4). The MLP (BFGS) model has very good performance in the two data sets (training and testing) with standard deviation of 15.030 and the 13.459, respectively (Table 5). The respective values of coefficient of determination ($R^2$) values and accuracy factor ($A_f$) for the two data sets are 0.929 and 1.420 for the training phase, and 0.811 and 1.210 for the testing phase (Table 4).

Fig. 4 presents a scatter plot of the MLP BFGS-simulated WQI values and residuals corresponding to the training, testing, and all data sets. The observed relationship between MLP BFGS-simulated WQI values and residuals for all the two shows complete independence and random distribution. It is further supported by the negligible small correlations ($R^2 = 0.000$ for training, $R^2 = 0.119$ for testing, and $R^2 = 0.014$ for all data sets). Fig. 6 explains that the points are well distributed on both sides of the horizontal line of zero ordinate representing the average of the residuals suggesting that the model fits the data well [46].

In order to identify the effect of input variables (water quality parameters) toward the output (WQI), the MLP BFGS neural network sensitivity analysis was calculated in both training and testing phases. Table 6 indicates that the fifth most effective water quality parameters for WQI decreasing, in descending order, are BOD₅, Cl⁻, NO₃, DCO, and TDS. The remaining water quality parameters according to their ranking in the testing phase are: oxygen saturation, phosphates, turbidity, pH, and water temperature. In light of these findings, the water quality monitoring agency may give priority consideration to these fifth water quality parameters.

The results of the MLP BFGS neural network and expert opinion (Table 7) are similar only in ranking the first, second, and third priority water quality parameters which are BOD₅, chloride, and NO₃ whilst they differ in ranking the remaining water quality parameters.

#### 3.3. Correlation matrix

The water temperature (WT) has significant environmental effects by influencing the physical, chemical, and biochemical. It was positively correlated with pH ($r = 0.61$) and oxygen saturation ($r = 0.62$) (Table 8). Table 8 shows a

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Data mean</th>
<th>Data SD</th>
<th>RMSE</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>43.590</td>
<td>15.030</td>
<td>0.009</td>
<td>0.975</td>
</tr>
<tr>
<td>Testing</td>
<td>53.343</td>
<td>13.459</td>
<td>0.007</td>
<td>0.911</td>
</tr>
</tbody>
</table>

### Table 4 Performance criteria in various MLP neural networks

<table>
<thead>
<tr>
<th>ANN</th>
<th>Architecture</th>
<th>Training data sets</th>
<th>Testing data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>$R^2$</td>
<td>$A_f$</td>
</tr>
<tr>
<td>MLP (CG 45)</td>
<td>0.033</td>
<td>0.814</td>
<td>1.595</td>
</tr>
<tr>
<td>MLP (CG 39)</td>
<td>0.021</td>
<td>0.895</td>
<td>1.571</td>
</tr>
<tr>
<td>MLP (BFGS 60)</td>
<td>0.009</td>
<td>0.929</td>
<td>1.420</td>
</tr>
</tbody>
</table>
significant and positive correlation between TDS, NO₃⁻, PO₄³⁻, BOD₅, COD, and Cl⁻ (r = 0.69–0.77), which are responsible for water contamination. The NO₃⁻ concentration showed a significant positive correlation with PO₄³⁻ (r = 0.88), BOD₅ (r = 0.78), COD (r = 0.69), and chloride (r = 0.82), and negative correlation with WQI (r = –0.79). PO₄³⁻ correlated reasonably well with BOD₅, COD, Chloride (r = 0.72–0.89), and WQI (r = –0.90) suggesting that PO₄³⁻ originated from anthropogenic sources. BOD₅ and COD are two parameters used to estimate the organic contamination load [47]. BOD₅ and COD showed a positive correlation between them, indicating contamination of organic matter. As also
shown in Table 8, both BOD$_5$ and COD present a significant positive correlation with chloride and a significant negative correlation with WQI.

3.4. Principal component analysis

PCA module is applied to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible the variability present in data set and to evaluate the relationship between variables, since they show the participation of individual chemicals in several influence factors [2].

Table 9 shows that there are 11 variables in the analysis, the number of principal components was chosen in accordance with Kaiser’s criterion, and Cattell’s scree test. It shows that the principal components with eigenvalues close to or greater than 1 were considered for interpretation. Therefore, two principal components were chosen for analysis with a cumulative variance of 73.28%. The remaining eigenvalues each account for less than 10% of the total variance.

The principal components loading of the different water quality parameters and WQI are presented in Table 9. The first principal component PC$_1$ explains 55.55% of the total variances and corresponds to the largest eigenvalue (6.11), PC$_1$ is highly correlated with TDS, nitrates, phosphates, BOD$_5$, COD, and chloride (negative correlation) and WQI (positive correlation). These are the parameters that primarily affect the Saf-Saf rivers’ water quality. The second principal component PC$_2$ corresponding to the second eigenvalue (1.94) accounts for 17.70% of the total variance.

3.5. Factor analysis

Factor analysis was carried out 11 variables to identify the various variifactors that influence each of them. Three variifactors VF$_i$s were obtained through FA performed on the PCs and it explaining more than 90% of the total variance (Fig. 8). The corresponding VF$_i$s, variable loadings are presented in Table 10. Varifactor 1, which explained 59.03% of the total variance, had strong negative loadings (=–0.94) on WQI, a positive loading on TDS, NO$_3$–, PO$_4^{3–}$, BOD$_5$, and COD. This varifactor can be interpreted as anthropogenic effects on surface water of Saf-Saf river basin. The projection of the cases on the factor plane (PC$_1$ × PC$_2$) shows that sampling sites were grouped into three main groups (Fig. 7b). The group I gathers the sampling sites which are typical by the average WQI and characterized by low values of chloride, TDS, NO$_3$–, PO$_4^{3–}$, BOD$_5$, and COD. The sampling sites of group I are located in upstream of Saf-Saf river basin and correspond to a relatively low pollution during two seasons (wet and dry season). The group II includes the sampling sites of Saf-Saf valley during dry season and the sampling sites which are located in downstream of Saf-Saf river basin during wet season. This group represents waters with marginal quality based on the WQI. The group III gathers the sampling sites (during dry season) located in downstream area which are characterized by the high values of chloride, TDS, NO$_3$–, PO$_4^{3–}$, BOD$_5$, and COD and very low WQI which showed evidence of surface water quality deterioration.

4. Conclusion

In this paper, we developed a new methodology based on a combination of artificial neural networks and multivariate
statistical techniques to forecast WQI of surface water in an unmonitored river basin. An MLP neural networks with three different algorithms were trained and tested using datasets (water quality parameters and WQI) measured during wet and dry season in 2015.

The predictive capability of the MLP model is determined using three criteria, namely, RMSE, coefficient of determination ($R^2$), and the accuracy factor ($A_f$). The results obtained in this paper show that MLP_BFGS neural network demonstrate to be the best ANN structure indicating that BOD, Chloride, NO$_3$, DCO, and TDS are the fifth most effective water quality parameters influencing WQI in Saf-Saf river. Selecting and ranking water quality parameters assist decision-makers and water managers to give a priority consideration to these fifth water quality parameters in terms of surface water monitoring.

Table 8
Correlation matrix – water quality parameters and WQI

<table>
<thead>
<tr>
<th></th>
<th>pH</th>
<th>WT</th>
<th>OS</th>
<th>TDS</th>
<th>Turbidity</th>
<th>NO$_3$</th>
<th>PO$_4$</th>
<th>BOD$_5$</th>
<th>COD</th>
<th>Cl$^-$</th>
<th>WQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT</td>
<td>0.61</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>0.27</td>
<td>0.62</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDS</td>
<td>-0.10</td>
<td>0.53</td>
<td>0.22</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turbidity</td>
<td>0.13</td>
<td>0.20</td>
<td>0.30</td>
<td>0.15</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO$_3$</td>
<td>0.17</td>
<td>0.48</td>
<td>0.35</td>
<td>0.69</td>
<td>0.12</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO$_4$</td>
<td>-0.04</td>
<td>0.38</td>
<td>0.19</td>
<td>0.77</td>
<td>0.15</td>
<td>0.88</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOD$_5$</td>
<td>-0.13</td>
<td>0.27</td>
<td>0.16</td>
<td>0.71</td>
<td>0.09</td>
<td>0.78</td>
<td>0.83</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COD</td>
<td>-0.09</td>
<td>0.30</td>
<td>0.12</td>
<td>0.67</td>
<td>0.13</td>
<td>0.69</td>
<td>0.72</td>
<td>0.93</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>-0.14</td>
<td>0.29</td>
<td>0.19</td>
<td>0.74</td>
<td>0.08</td>
<td>0.82</td>
<td>0.89</td>
<td>0.87</td>
<td>0.78</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>WQI</td>
<td>0.09</td>
<td>-0.42</td>
<td>-0.12</td>
<td>-0.86</td>
<td>-0.19</td>
<td>-0.79</td>
<td>-0.90</td>
<td>-0.84</td>
<td>-0.78</td>
<td>-0.87</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Underlined correlations are significant at $p < 0.0500$

Table 9
Loadings of water quality parameters (10) and WQI on principal components for the whole datasets (Underlined loadings are $>0.70$)

<table>
<thead>
<tr>
<th>PC$_1$</th>
<th>pH</th>
<th>WT</th>
<th>OS</th>
<th>TDS</th>
<th>Turbidity</th>
<th>NO$_3$</th>
<th>PO$_4$</th>
<th>BOD$_5$</th>
<th>COD</th>
<th>Cl$^-$</th>
<th>WQI</th>
<th>Eigenvalue</th>
<th>% variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC$_1$</td>
<td>-0.01</td>
<td>-0.51</td>
<td>-0.30</td>
<td>-0.86</td>
<td>-0.18</td>
<td>-0.90</td>
<td>-0.93</td>
<td>-0.92</td>
<td>-0.86</td>
<td>-0.92</td>
<td>0.94</td>
<td>6.11</td>
<td>55.55</td>
</tr>
<tr>
<td>PC$_2$</td>
<td>-0.84</td>
<td>-0.78</td>
<td>-0.65</td>
<td>0.03</td>
<td>-0.20</td>
<td>-0.13</td>
<td>0.09</td>
<td>0.21</td>
<td>0.18</td>
<td>0.19</td>
<td>-0.12</td>
<td>1.94</td>
<td>17.70</td>
</tr>
</tbody>
</table>

Fig. 7. Projection of variables and sampling sites on the factor-plane (a and b) PC$_1 \times$ PC$_2$
PCA combined with FA were used to assess variations in surface water quality of Saf-Saf river basin, both in time and space. It shows three groups of sampling sites, the group I on the right side of the PC1 gathers upstream sampling sites during dry season; it is characterized by very low WQI and it corresponds to high concentration in BOD5, Chloride, DCO, and TDS reflecting very polluted surface water.

Table 10
Factor loadings – water quality parameters and WQI (rotation: Varimax normalized) extraction: principal components (underlined loadings are >0.70)

<table>
<thead>
<tr>
<th>VF1</th>
<th>pH</th>
<th>WT</th>
<th>OS</th>
<th>TDS</th>
<th>Turbidity</th>
<th>NO3</th>
<th>PO4</th>
<th>BOD5</th>
<th>COD</th>
<th>Cl-</th>
<th>WQI</th>
<th>% variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>VF1</td>
<td>-0.18</td>
<td>0.32</td>
<td>0.16</td>
<td>0.85</td>
<td>0.11</td>
<td>0.85</td>
<td>0.93</td>
<td>0.94</td>
<td>0.88</td>
<td>0.94</td>
<td>-0.94</td>
<td>59.03</td>
</tr>
<tr>
<td>VF2</td>
<td>0.79</td>
<td>0.86</td>
<td>0.76</td>
<td>0.15</td>
<td>0.07</td>
<td>0.33</td>
<td>0.11</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>-0.07</td>
<td>21.07</td>
</tr>
<tr>
<td>VF3</td>
<td>0.23</td>
<td>0.14</td>
<td>-0.25</td>
<td>0.07</td>
<td>0.95</td>
<td>0.01</td>
<td>0.06</td>
<td>-0.02</td>
<td>0.05</td>
<td>-0.04</td>
<td>-0.13</td>
<td>10.62</td>
</tr>
</tbody>
</table>

Fig. 8. Factor loadings, VF1 vs. VF2 vs. VF3 – water quality parameters and WQI (rotation: Varimax normalized).

References