Journal of Geoscience and Environment Protection, 2017, 5, 99-124
http://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Application of ANN and MLR Models on Groundwater Quality Using CWQI at Lawspet, Puducherry in India

N. Suresh Nathan, R. Saravanane, T. Sundararajan

Department of Civil Engineering, Pondicherry Engineering College, Puducherry, India
Email: suresh_eepdy@yahoo.com

How to cite this paper: Nathan, N.S., Saravanane, R. and Sundararajan, T. (2017) Application of ANN and MLR Models on Groundwater Quality Using CWQI at Lawspet, Puducherry in India. *Journal of Geoscience and Environment Protection*, **5**, 99-124.
https://doi.org/10.4236/gep.2017.53008

Received: January 4, 2017
Accepted: March 13, 2017
Published: March 16, 2017

Abstract

With respect to groundwater deterioration from human activities a unique situation of co-disposal of non-engineered Municipal Solid Waste (MSW) dumping and Secondary Wastewater (SWW) disposal on land prevails simultaneously within the same campus at Puducherry in India. Broadly the objective of the study is to apply and compare Artificial Neural Network (ANN) and Multi Linear Regression (MLR) models on groundwater quality applying Canadian Water Quality Index (CWQI). Totally, 1065 water samples from 68 bore wells were collected for two years on monthly basis and tested for 17 physio-chemical and bacteriological parameters. However the study was restricted to the pollution aspects of 10 physio-chemical parameters such as EC, TDS, TH, HCO_3^-, Cl^-, SO_4^{2-}, Na^+, Ca^{2+}, Mg^{2+} and K^+. As there is wide spatial variation (2 to 3 km radius) with ground elevation (more than 45 m) among the bore wells it is appropriate to study the groundwater quality using Multivariate Statistical Analysis and ANN. The selected ten parameters were subjected to Hierarchical Cluster Analysis (HCA) and the clustering procedure generated three well defined clusters. Cluster wise important physio-chemical attributes which were altered by MSW and SWW operations, are statistically assessed. The CWQI was evolved with the objective to deliver a mechanism for interpreting the water quality data for all three clusters. The ANOVA test results *viz.*, F-statistic ($F = 134.55$) and p-value ($p = 0.000 < 0.05$) showed that there are significant changes in the average values of CWQI among the three clusters, thereby confirming the formation of clusters due to anthropogenic activities. The CWQI simulation was performed using MLR and ANN models for all three clusters. Totally, 1 MLR and 9 ANN models were considered for simulation. Further the performances of ten models were compared using R^2, RMSE and MAE (quantitative indicators). The analyses of the results revealed that both MLR and ANN models were fairly good in pre-

dicting the CWQI in Clusters 1 and 2 with high R^2, low RMSE and MAE values but in Cluster 3 only ANN model fared well. Thus this study will be very useful to decision makers in solving water quality problems.

Keywords

Canadian Water Quality Index, Multi-Linear Regression, Artificial Neural Network, Simulation, Comparison

1. Introduction

Fresh water covers about 2.5% of earth's water out of which groundwater constitutes about 30.1% (https://water.usgs.gov). Even though groundwater is abundant, it may still be unusable when its quality is considerably deteriorated by chemical and bacteriological contamination due to anthropogenic activities in social and industrial sectors. The increase in population and urbanization plays a pressing role in augmenting the demand for water supply in municipal and industrial sectors.

The natural processes like weathering of rocks/soils, atmospheric precipitation etc., play an important role in the chemical constitution of groundwater [1] [2]. Further exploitation of groundwater due to agrarian, industrial and urban activities plays a critical part in the degradation of groundwater quality [3]. However in the recent years, anthropogenic activities like discharge of untreated or partially treated waste water, mining and related activities, solid waste dumping, contaminated agricultural runoff due to pesticides etc., compound the likelihood of groundwater deterioration [4].

Furthermore the groundwater qualitatively relies on the physio-chemical and bacteriological quality of recharged water, inland surface run off and subterranean geochemical responses. Cyclic changes in groundwater may also be brought about by hydrological and anthropogenic components [5] [6].

Casual regulation and multiplying anthropogenic activities frequently lead to the disproportionate dispensation of chemical components in groundwater, resulting in varying analytical data. So there is a critical demand for the planners, administrators and managers to distinguish the groundwater pollution and search for a fruitful and reliable system for regulating ground water resources and allied pollution [6].

The monitoring and analysis of water quality status are absolutely necessary to detect long-term trends in selected water quality parameters so as to discern prospective water quality problems. It is intended to determine the most contributing parameters which cause alterations in groundwater quality by calculating water quality indices for various water uses like drinking water, irrigation, recreation, aquatic life etc. The water quality assessment shows the variation of water quality parameters. At the same moment, good quality of water must be adequately accessible to nurture hale and healthy life.

Regionally because of non-availability of surface water, the entire population of Puducherry, India has to rely upon groundwater reserves. With respect to groundwater deterioration from human activities at Puducherry, we then have two significant aspects:

- Contamination based on non-engineered Municipal Solid Waste (MSW) dumping
- Partially treated or Secondary Wastewater (SWW) application on land

In order to assess the groundwater qualitatively, dependable data on water quality is required, which can be acquired through routine water quality surveillance programs. These programs usually generate huge and complicated data matrix containing a number of water quality attributes, which are generally hard to comprehend and evaluate the water quality as a whole. To overcome this, a mathematical technique, which transforms the massive of water quality data into a single count, such as WQI is required to ascertain the extent of pollution in water bodies. Thus WQI is a powerful tool to get overall information on water quality in a readily explicit form that can be utilized by administrators, decision makers and people. The theory of WQI is contingent on the precept of collating water quality parameters with respect to controlling threshold limits.

A single WQI value gives information more precisely and it is easy to understand than a long list of parametric values. Additionally, WQI also facilitates comparison between different sampling locations at different points of time. Considering the simplicity and reliable approach of WQI, it is ascertained that these indices will furnish explicit outlines of overall water quality and possible drifts. Thus, WQI can be used to furnish a comprehensive summary of environmental performance that can be expressed to the public in an understandable pattern. While appreciating the importance and usability of WQI, it is important to know about the limitations of WQI:

1) Lack of information due to amalgamation of several parameters to a single index measure.

2) Sensitivity of the results due to the formulation of the index.

3) Loss of information due to exchange among parameters, and

4) Want of flexibility of the index to divergent ecosystems.

Thus, WQI can be used as a powerful tool to get overall information and a comprehensive summary of environmental performance on water quality in a readily explicit form that can be utilized by administrators, decision makers and people in an understandable pattern [7]. Further, WQI is neither a replacement to the detailed analysis of environmental monitoring and modeling, nor should it be the only tool for the management of water bodies.

In this situation, if deterministic models using MLR or simulation models like ANN could be developed for finding out water quality index, then it will be of great help to the managerial community to closely monitor the groundwater quality and the models so developed, could be a reliable alternative to the complicated and time consuming water quality index calculations. Further these models are very practical, robust and cost effective.

2. Study Area and Present Scenario

Puducherry is a Union Territory in India with an extent of 293 km^2. The entire urban and sub urban areas of Puducherry are divided into nine zones for the purpose of water supply and comprehensive underground drainage system. Among the nine zones, Zone V (Lawspet) is a likely zone for groundwater exploitation. The borewells in Zone V are the only sources of water supply to coastal zones like Zone II (Muthialpet) and Zone V (Lawspet), as the current water supply in these zones are contaminated due to sea water intrusion. Of late, Zone V area is also getting affected due to the above said anthropogenic activities. Under these circumstances, it was decided to adopt Zone V which is a potential groundwater source, for study purposes in order to prevent further contamination.

The study area falls in Zone V (Lawspet) area, wherein the STP and solid waste landfill are located in the same campus at Karuvadikuppam, Lawspet at Latitude 11˚58'16"N and Longitude 79˚48'11"E on the northern part of Puducherry, India (**Figure 1**). The terrain declines from North to South and the ground elevation ranges from 53 m to 6 m as shown in **Figure 1**. The area is identified with tropical climate with a mean yearly precipitation of 1200 mm, 35% of which takes place during the South-West monsoon from June to September and the remaining 65% befalls during the North-East monsoon i.e. from October to December [8]. Presently 15 MLD of wastewater is treated using four serially connected facultative oxidation ponds and 1 UASB of capacity 2.5 MLD. Domestic sewage of BOD 250 mg/L is treated with a removal efficiency of 65%. Nearly, 12.5 MLD of partially treated SWW is discharged into a recharge pond area of 18 acres, since 1980 [8].

A portion of Sewage Treatment Plant (STP) site at Karuvadikuppam is used as solid waste landfill. Solid waste tipping started in 2004 and discontinued in 2013 only and it spreads over an area of 21 acres approximately [9]. It is a non-engineered low lying open dump. The land fill is unlined and the solid waste has been dumped indiscriminately in an unscientific way and irregular fashion. The solid waste landfill height varies from 2 m to 6 m. So in Puducherry, a unique situation of co-disposal of MSW dumping and SWW disposal on land prevails simultaneously within the same campus.

Against this back drop an effort has been attempted to investigate the spatial variation of water quality using, a readily understandable indicator, *i.e.* water quality index (WQI) for various intended uses (drinking and domestic uses).

Broadly the objectives of the study were to 1) to apply the water quality index (WQI) so that the changes in the ground water quality of the study area can be monitored mainly for drinking purposes 2) to develop a Multivariate Linear Regression (MLR) model, to simulate WQI which is commonly used as an indicator of groundwater pollution 3) to establish an Artificial Neural Network (ANN) model that can be applied to directly foretell the water quality condition in the study area and thus furnish a dependable substitute to the WQI calculation method presently in use and 4) to evaluate the effectiveness of two data oriented methodologies *viz.*, MLR and ANN.

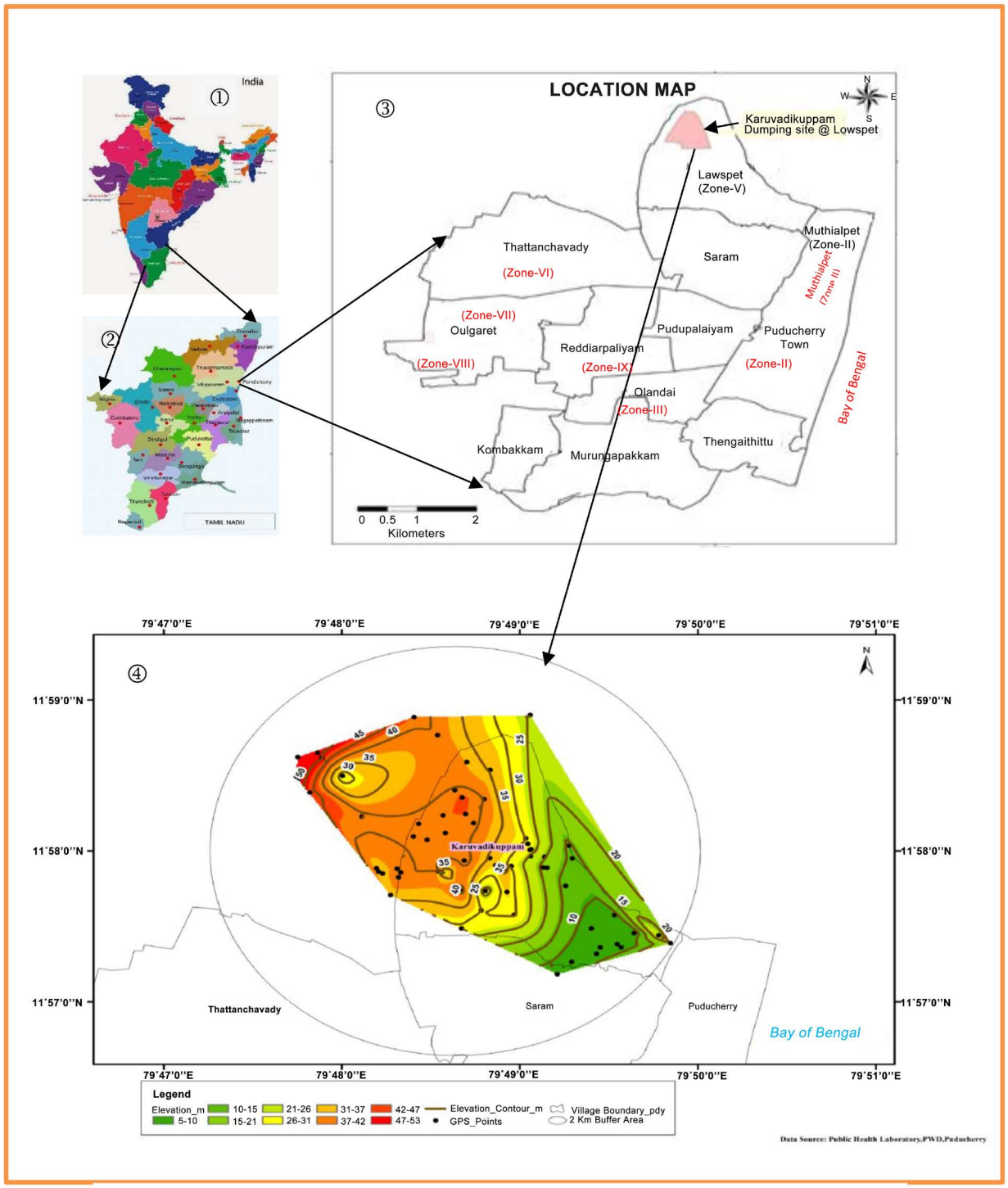

Figure 1. Location map, elevation and sampling borewell sites in study area.

3. Methodology

3.1. Sample Collection and Monitoring of Borewells

Nearly 125 water supply and agricultural borewells are located in and around STP within a radial distance of 2.5 km from STP and solid waste landfill. To accurately represent the groundwater quality, a sampling strategy was formulated

to include a wide range of bore wells at the pivotal locations. Totally, 68 bore-wells were identified in and around the study area and depicted in **Figure 1**. All the bore wells were considered for investigation and water samples were collected every month from Jan 2014-Dec 2015 from solid waste dump area, recharge pond area, sewage farm area (existing) and peripheral area (private & Govt.) in order to study the seasonal and spatial variations.

3.2. Physio-Chemical Analysis of Groundwater

Water samples were collected from the borewells after pumping for 15 minutes. The samples were analyzed in the Public Health Laboratory, PWD, Puducherry, India. Totally 1065 water samples were collected and tested for 17 physiochemical and bacteriological parameters viz. EC, pH, TDS, Alkalinity, HCO_3^-, TH, Ca^{2+}, Mg^{2+}, Fe^{2+}, Cl^-, SO_4^{2-}, NO_3^-, Na^+, F^-, K^+, PO_4^{3-}, Si^{4+}, BOD, COD. Total Coliforms and Faecal Coliforms according to the standard methods [9] [10]. However the study was restricted to the pollution aspects of tenwater quality parameters viz. EC, TDS, HCO_3^-, TH, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, Na^+, and K^+.

3.3. Canadian Water Quality Index (CWQI)

The Canadian Council of Ministers of the Environment Water Quality Index (CWQI) is a well-established and universally accepted model for evaluating the WQI. The CWQI compares observations to a guideline value, which can be a water quality criterion or a locality specific chemical composition of a hydro geological parameter.

The CWQI was formulated with the purpose of rendering a mechanism for reducing the documentation of water quality data [11] [12] [13]. As a compendious tool, it furnishes a wide sketch of water quality data and it is functional for various purposes including drinking water quality, data communications, ambient water quality data processing, combined watershed designing, management and policy decisions in the water supply sector.

The model essentially consists of three measures of variance based on selected guideline values or threshold limits (scope, frequency, amplitude) as detailed below:

1) The number of variables whose objectives are not met (scope).

2) The frequency with which the objectives are not met (frequency).

3) The amount by which the objectives are not met (amplitude).

These values are concerted to bring about a single value (between 0 and 100) portraying the water quality. A value of 100 is the best possible index score and a value of 0 is the worst possible. The Canadian Water Quality Index (CWQI) is computed as follows:

$$F_1(Scope) = \frac{\text{Numberoffailedvariables}}{\text{Totalnumberofvariables}} \times 100$$

$$F_2(\text{Frequency}) = \frac{\text{Numberoffailedtests}}{\text{Totalnumberoftests}} \times 100$$

$$F_3\left(\text{Amplitude}\right)=\frac{nse}{0.01nse+0.01}$$

$$nse=\sum_{i=1}^{n}\left(\text{excursions}_i\right)/\text{No.oftests}$$

$$\text{excursions}_i=\frac{\text{Failedtestvalue}}{\text{Objective}_i}-1$$

$$\text{CWQI}=100-\frac{\sqrt{F_1^2+F_2^2+F_3^2}}{1.732}$$

A sample calculation for computing CWQI (refer Table 1) for borewell (BW1) is as follows:

$$F_1=100\times 8/10=80;\ F_2=100\times 8/10=80$$

$$\text{excursion}=\left(1423/750+905/500+\cdots\right)-8=5.113;\ nse=5.113/10=0.511$$

$$F_3=nse/\left(\left(0.01\right)\left(nse\right)+0.01\right)=33.83$$

$$CWQI=100-\frac{\sqrt{F_1^2+F_2^2+F_3^2}}{1.732}=31.82\text{say}32$$

3.4. Multivariate Statistical Analysis

As there is wide spatial variation (2 to 3 km radius) with ground elevation (more than 45 m) among the borewells it is appropriate to investigate the quality of the ground water in the entire study area applying multivariate statistical analysis.

Uncertainty is innate in all methods of evaluating groundwater pollution, arising from missing data, the natural, spatial and temporal in consistency of the hydrogeological variables in the field, and in mathematical computation. Conventional procedures are inferior at addressing the non-linearity, subjectivity, and intricacy of the cause-effect association between water quality parameters and conditions but still they are the presently accepted procedures.

In similar situations, multivariate statistical methods [14] [15] [16] [17] and artificial neural networks, can be effectively used in a wide variety of environmental applications. The results demonstrated that the combined approach effectively interpreted the geological significance of the factors, and also reduced the area of exploration targets.

3.5. Artificial Neural Networks (ANN)

Several deterministic models have been tried in the past for prediction of CWQI

Table 1. Mean physio-chemical test results of BW1.

	EC	TDS	TH	HCO_3^-	Cl^-	SO_4^{2-}	Na^+	Ca^{2+}	Mg^{2+}	K^+
BW1	1423	905	377	209	323	85	178	83	34	5
Guideline Value	750	500	300	200	250	200	50	75	30	10

Note: EC-μs/cm, all other parameters mg/L.

in groundwater. These models require input data, model parameters, and extensive information to obtain results. But, in practice the statistical precision of the models is not encouraging because natural systems tend to be too complex for deterministic modeling [18] [19].

Further, because of large number of factors affecting the water quality, and their complicated nonlinear relationships with the variables, the traditional deterministic models are not easy to handle. On the contrary ANNs provide a quick, flexible and reliable means of creating models for estimating groundwater quality. Currently ANNs have revealed very good realization as regression tools, chiefly when applied for pattern recognition and function estimation. Thus ANN is used as an approximation tool rather than a complex mathematical calculation, which results in admissible deviation of predicted value from observed data [20] [21].

In relation to the conventional approaches, ANNs admit approximate or missing data, inexact results, and they are less susceptible to outliers. Further they are highly parallel, *i.e.*, their multitudinous independent operations can be handled concurrently. Because of parallel processing architecture, ANN is competent enough to manipulate complicated numerations, thus making it the most popular technique today for high speed computing of large data. In addition, there are many advantages in problem solving as elaborated below:

1) Application of a neural network does not call for previous comprehension of the underlying process, so it can be employed to solve the problems vaguely described.

2) No need to identify all the complex associations among various features of the process under analysis.

3) A conventional optimisation procedure or statistical model delivers a solution only when executed completely whereas a neural network always converges to a local optimal result.

4) The model has more forbearance to noise and ambiguous data thereby requiring less information for model development.

5) The findings are the outcome of the generic behaviour of data, as such the influence of outlier is reduced.

For these reasons ANNs are found to be more suitable for handling various hydrological modeling problems [22] [23].

3.5.1. Structure of an ANN

ANN is a simulation of the real nervous system in other words, it is a numerical model contingent on biological neural networks. It is a system which consists of a collection of units called "neurons" communicating with each other in a network that works to produce a simulated output. ANNs are inspired by the activity of human brain. The basic units of any biological neural system are neurons, which are classified into sets, consisting of millions of them, organized in layers and constitute their own functional arrangement. A set of these subsystems create a global system [24] [25].

3.5.2. Components of an ANN

Typically a neuron receives many parallel and multiple inputs. Each input has its own relative weight which signifies the importance of the input within the activation function of the neuron. These weights do the same role played by the biological neurons in synapses. In both cases, some inputs are more important than others so they have more involvement in the processing of the neuron and to produce a neuronal result. The weights are coefficients that can be adapted within the network depending on the intensity of the input signal, received by the artificial neuron [26].

Based on the inputs and weights, the summing part provides the potential postsynaptic value "h_i" of the neuron. The most common function is the sum of all weights and inputs, by grouping the inputs and weights in two vectors ($x_1, x_2, \cdots, x_n$) and ($w_{1j}, w_{2j}, \cdots, w_{nj}$) and then calculate this amount making the scalar product of two vectors.

$$h_i(t) = Sw_{ij} * x_i \tag{1}$$

where $h_i(t)$ is post synaptic potential.

The inputs and weights can be combined in different ways before transferring the value to the activation function. The specific algorithm for the propagation of neural inputs depends on the choice of architecture. The result of the summing part in most cases is a weighed sum, which is transformed into the actual output of the neuron through an algorithmic process known as activation function.

$$a_i(t) = f_i(a_i(t-1) * h_i(t) \tag{2}$$

The activation function depends on the postsynaptic potential "$h_i(t)$" and its previous state of activation. However, in many models of ANN, the current state of the neuron does not depend on its previous state "$a_i(t-1)$", but only on the current state.

$$a_i(t) = f_i(h_i(t)) \tag{3}$$

In the activation function, the value of the output combination can be compared with a threshold value for determining the output of the neuron. If the sum is greater than the threshold value, a neuron signal is generated. If the sum is less than the threshold, no signal is generated. Usually the threshold value, or transfer function value is typically nonlinear.

Before applying the activation function, some noise is added to the inputs. The source and amount of this noise are determined by training of a particular network. This noise is commonly known as temperature of the neuron. In fact by adding different noise levels to the result of the combination or summing, a model more similar to the brain can be created.

3.5.3. Activation Function

The activation (transfer) function establishes the reaction of a node to the total input signal it acquires. Generally hidden layer utilizes logistic transfer function. By means of an activation function, from hidden layer to output layer, a linear

transfer function is applied. Most commonly used non-linear sigmoid function is

$$f(s) = 1/\left(1+e^{-as}\right)$$

Also hyperbolic tangent transfer function is used in many networks as follows:

$$f(s) = \tanh(s) = \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}} = \frac{1-e^{-2s}}{1+e^{-2s}}$$

$$\text{where} s_i = \sum_{i=1}^{n} w_i x_i$$

w_i—weights & x_i—input variables.

3.5.4. Architecture of an ANN

Generally ANN models (**Figure 2**) were specified by the network topology, training and/or learning rules [27] [28]. These ambiences have primarily configured the network behaviour with three different layers in the network topology which can be distinguished as:

1) An input layer: contains neurons that acquire information from the environment and connects the input information to the system (network).

2) Hidden layer: acting as an intermediate computational layer.

3) Output layer: the neurons provide the response and produce the desired output of the neural network.

The connections between neurons can be excitatory or inhibitory: a negative synaptic weight defines a negative inhibitory connection, while a positive determines excitatory connection. Intra-layer connections, also called side connections, take place between neurons in one single layer, while the inter-layer connections occur between neurons in different layers. There are also feedback

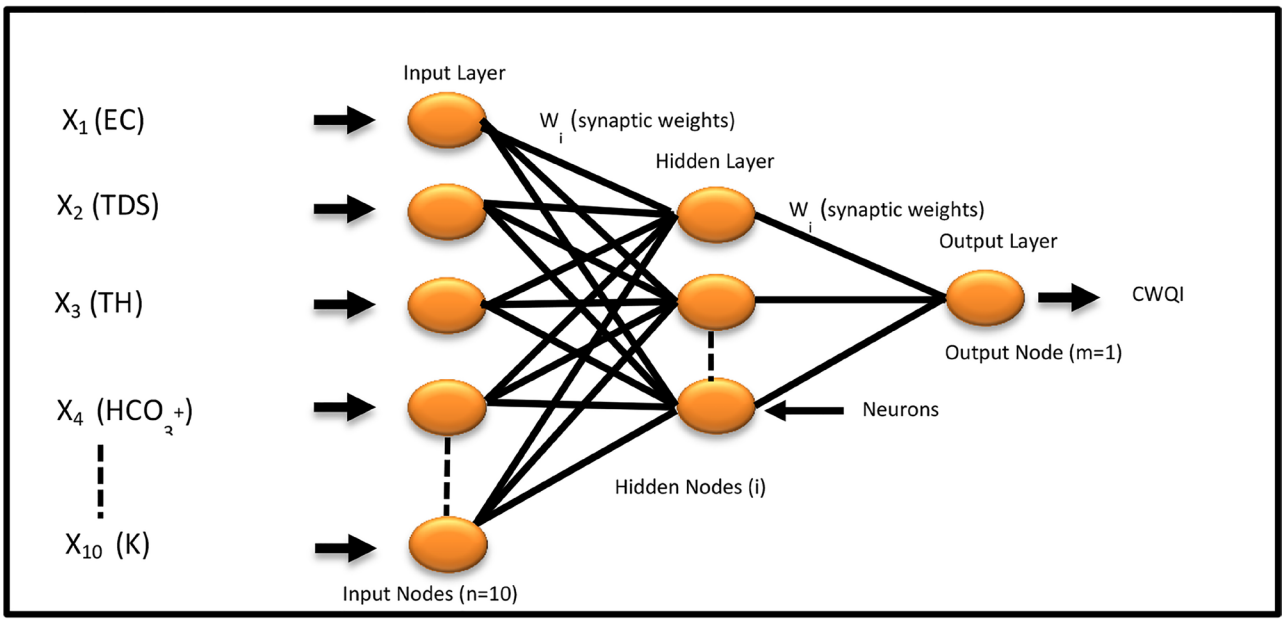

Figure 2. Architecture of an ANN model.

connections that have an opposite way input-output. Based on these concepts, different neural architectures can set:

- Single Layer Networks consist just one layer of neurons.
- Multi-Layer Networks are those whose neurons are organized in several layers, in response to the data flow in a neural network.
- Feed Forward Networks circulate the information unidirectionally from the input neurons to the output neurons.
- Feedback Networks circulate the information between the layers in any direction.

3.5.5. Training and Testing Algorithm

In this research work, Multilayer Perceptron (MLP), a feed forward kind of ANN model is employed. In this model, a set of input data is fed into a network to receive a set of suitable output data after due mathematical processing. The MLP model is a network comprises of multiple layers of nodes (neurons) and these layers are interconnected from one layer to another. All the interconnected nodes of various layers form a directed graphical network system. The hidden layer and the output layer are connected through neurons, exercising a nonlinear activation function by a technique known as "back propagation" to train the network system. To standardize the model, the entire data set is segregated into three phases. The first phase is the learning phase which is utilized to train the network. The goal of training is to guarantee that the network replicates the inherent characteristic of the information availed in the ANN modelling. ANN weights and biases are fixed during the training procedure. The input variables and already ascertained output parameters decide the associated weights in such a way that the predicted and observed values are in agreement. Secondly the ANN models are put to testing in order to fix up the stopping criterion as when to stop. Lastly the model is validated utilizing the data which are not included in the training phase [29] [30] [31].

3.5.6. Model Performance Appraisal

Performance appraisal of the MLR and ANN models is accomplished using three statistical indices, namely: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) to study the capability of simulating cluster wise CWQI. MAE demonstrates the mean of the errors simulated by the developed model and is employed to detect the proximity of simulated values (observed values). RMSE signifies the comprehensive variation between observed and simulated values. R^2 represents the measure of the total variance with respect to the observed values, which can be explained by the developed model. The mathematical expressions for MAE and RMSE are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_{oi} - x_{pi}\right|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{oi} - x_{pi}\right)^2}$$

where x_{oi}—observed CWQI and x_p—predicted CWQI.

4. Results and Discussion

4.1. Hierarchical Cluster Analysis (HCA)

HCA visualizes intra-relationship among the parameters for a good perception of the studied system. Divergent sampling locations in the study area can be grouped into clusters to spatially explain the similarity in chemical composition of the groundwater quality among the bore wells [32] [33] [34]. The selected ten hydro-chemical parameters in the study area were subjected to HCA. The clustering procedure generated three well defined clusters. Cluster 1 involves 28 bore wells, forming 41% of the sampling stations. Cluster 2 comprises of 8 bore wells, representing 12% of the sampling stations. Cluster 3 accounts for 32 bore wells comprising of 47% of the sampling stations. Clusters 1, 2 and 3 correspond to polluted, highly polluted and non-polluted regions of the study area.

4.2. Descriptive Statistics

During this study some important physio-chemical attributes from shallow groundwater in the study area were obtained and measured. The main intention of this study was to evaluate, determine, predict and compare the groundwater quality dispensation and extent of prospective contamination in the study area using MLR and ANN models. Besides statistically reporting the current status of shallow groundwater for impending comparisons, the study will also be precious to the managers answerable for groundwater development, regulation, exploitation and deterioration. Cluster wise important physio-chemical attributes (EC, TH, HCO_3^-, Cl^-, SO_4^{2-}, Na^+, Ca^{2+},Mg^{2+} and K^+) of water quality impaired by MSW & SWW exercises, apprehending the complete geology and ambient condition, are statistically assessed and presented in Tables 2-4.

Table 2. Descriptive statistics Cluster 1.

Statistic	EC	TDS	TH	HCO_3^-	Cl^-	SO_4^{2-}	Na^+	Ca^{2+}	Mg^{2+}	K^+	CWQI-OBS
No. of observations	28	28	28	28	28	28	28	28	28	28	28
Minimum	1004.0	634.0	226.0	147.2	194.9	50.3	73.2	54.3	20.2	2.6	31.2
Maximum	1626.0	1023.0	464.0	389.5	385.2	85.3	216.0	108.9	46.0	6.0	73.5
Range	622.0	389.0	238.0	242.4	190.3	35.0	142.8	54.6	25.8	3.4	42.3
1st Quartile	1123.0	695.3	247.3	204.5	228.0	62.2	148.1	59.9	25.7	3.5	39.5
Median	1238.0	781.5	273.0	234.6	261.8	65.6	166.4	65.6	31.4	3.9	48.5
3rd Quartile	1388.5	861.5	319.8	276.5	305.9	69.7	190.5	71.4	37.0	4.3	64.9
Mean	1253.5	787.3	290.4	242.7	267.4	66.4	165.1	67.3	32.1	4.0	51.6
Variance (n)	29074.6	12346.4	2864.1	3630.4	2434.5	48.9	1008.8	120.0	48.4	0.5	175.0
Standard deviation (n)	170.5	111.1	53.5	60.3	49.3	7.0	31.8	11.0	7.0	0.7	13.2
Skewness (Pearson)	0.4	0.4	1.3	0.6	0.6	0.4	−0.7	1.9	0.3	0.7	0.3
Kurtosis (Pearson)	−0.8	−0.8	1.8	−0.2	−0.5	0.8	0.5	5.0	−0.8	0.5	−1.2
Standard error of the mean	32.8	21.4	10.3	11.6	9.5	1.3	6.1	2.1	1.3	0.1	2.5

NOTE: EC-µs/cm, all other parameters mg/L.

Table 3. Descriptive statistics Cluster 2.

Statistic	EC	TDS	TH	HCO_3^-	Cl^-	SO_4^{2-}	Na^+	Ca^{2+}	Mg^{2+}	K^+	CWQI-OBS
No. of observations	8	8	8	8	8	8	8	8	8	8	8
Minimum	1797.0	1134.0	357.0	276.7	428.9	60.3	224.2	63.0	39.5	3.8	27.7
Maximum	2176.0	1356.0	453.0	458.7	517.1	126.3	317.9	89.5	67.5	10.6	35.9
Range	379.0	222.0	96.0	182.0	88.2	66.0	93.7	26.5	28.1	6.7	8.2
1st Quartile	1936.5	1225.0	381.0	353.5	448.2	75.2	275.4	74.0	45.0	6.5	28.2
Median	2035.5	1274.5	399.0	392.5	463.0	89.5	279.1	85.0	52.5	8.2	29.1
3rd Quartile	2108.8	1338.5	423.0	421.9	489.9	97.9	288.4	87.3	54.2	9.0	34.5
Mean	2010.4	1266.4	404.0	381.3	468.1	89.4	278.8	80.6	51.0	7.6	30.9
Variance (n)	16243.7	5951.2	1025.0	3242.6	834.2	392.1	607.6	76.4	73.5	5.6	10.8
Standard deviation (n)	127.5	77.1	32.0	56.9	28.9	19.8	24.7	8.7	8.6	2.4	3.3
Skewness (Pearson)	−0.5	−0.5	0.2	−0.5	0.3	0.4	−0.8	−0.9	0.3	−0.5	0.5
Kurtosis (Pearson)	−1.1	−1.1	−1.2	−0.9	−1.1	−0.7	0.8	−0.6	−0.5	−1.0	−1.6
Standard error of the mean	48.2	29.2	12.1	21.5	10.9	7.5	9.3	3.3	3.2	0.9	1.2

NOTE: EC-μs/cm, all other parameters mg/L.

Table 4. Descriptive statistics Cluster 3.

Statistic	EC	TDS	TH	HCO_3^-	Cl^-	SO_4^{2-}	Na^+	Ca^{2+}	Mg^{2+}	K^+	CWQI-OBS
No. of observations	32	32	32	32	32	32	32	32	32	32	32
Minimum	155.0	97.0	60.0	44.8	17.3	6.9	13.4	15.9	5.6	1.4	58.2
Maximum	848.0	548.0	360.0	436.4	190.2	101.0	115.2	80.1	41.0	4.3	91.8
Range	693.0	451.0	300.0	391.6	172.8	94.1	101.8	64.2	35.4	2.9	33.6
1st Quartile	327.5	209.0	116.5	77.1	29.6	17.9	23.9	31.1	9.2	1.8	82.9
Median	417.0	265.5	144.5	111.4	39.9	24.4	32.0	40.1	15.1	2.0	84.6
3rd Quartile	588.5	379.3	183.8	191.7	57.1	39.9	46.0	51.9	17.4	2.9	89.1
Mean	461.0	292.8	162.0	154.2	52.6	32.5	40.3	42.1	15.4	2.3	84.6
Variance (n)	35013.1	14187.8	5015.3	10825.9	1507.2	485.4	569.7	239.2	65.8	0.6	44.8
Standard deviation (n)	187.1	119.1	70.8	104.0	38.8	22.0	23.9	15.5	8.1	0.8	6.7
Skewness (Pearson)	0.5	0.5	1.1	1.2	2.2	1.4	1.4	0.5	1.3	0.8	−2.0
Kurtosis (Pearson)	−0.8	−0.7	0.8	0.3	4.4	1.4	1.5	−0.2	1.4	−0.4	5.4
Standard error of the mean	33.6	21.4	12.7	18.7	7.0	4.0	4.3	2.8	1.5	0.1	1.2

NOTE: EC-μs/cm, all other parameters mg/L.

4.3. Canadian Water Quality Index (CWQI)

CWQI has been formulated, based on the conceptual framework of Canadian Council of Ministers of the Environment [35] [36] [37]. CWQI so developed reflected the physio-chemical quality of the groundwater in the study area. The results from the model are an evidence of the degrading nature of groundwater in the borewell locations. The cluster wise Descriptive Statistics of CWQI values are presented in Tables 2-4. From the findings, it can be observed that the mean value of CWQI in Cluster 1 is 51.67 and the groundwater quality falls under the

category "marginal". Similarly the mean values of CWQI for Clusters 2 and 3 are 30.90 and 84.60 respectively and based on this, the overall groundwater quality of Clusters 2 and 3 can be ascribed as "poor" and "good". The poor nature of groundwater in Clusters 1 and 2 is mainly attributed to anthropogenic activities *viz.* 1) indiscriminate solid waste dumping and 2) partially treated SWW land application.

4.3.1. Analysis of Variance (ANOVA)

ANOVA has wide applicability in groundwater quality problems as a versatile diagnostic tool. The parametric one way ANOVA is an addendum of the t-test to multiple sample groups. ANOVA tests for significant differences in one or more clusters. If an overall significant difference is found as measured by F-statistic, post-hoc statistical contrasts may be used to determine where the differences lie among individual group means. In the CWQI monitoring context, only differences of mean relative to background are considered to be important.

The one way ANOVA technique generates one way analysis of variance for a quantitative dependent variable by a single factor independent variable. The purpose of using ANOVA is to address the following questions:

1) What are the main effects of independent variables (monitoring locations/clusters) on dependent variables (*i.e.* mean value of CWQI)?

2) What are the interactions among the independent variables?

Thus, one way ANOVA identifies spatial variability among monitoring bore-wells. Equality of variances among the clusters is evaluated with ANOVA and if it identifies significant differences then natural spatial variability is the likely cause. ANOVA compares the average values of CWQI among clusters to determine whether they are from same continuous distribution and whether significant differences existed between the mean values of CWQI among the clusters.

The hypothesis used is as follows:

H_0: There is no difference in the average levels of CWQI between the Clusters 1, 2 and 3.

H_1: There are differences in the average levels of CWQI between the Clusters 1, 2 and 3.

Decision making is the rejection of H_o if P value is less than α. The F-statistic and Sig (significance) conform the differentiation of clusters. $P < 0.05$ shows that high variations of CWQI in terms of their spatial distribution in the study area and consequently it may be concluded that H_0 is rejected and H_1 is accepted. In other words there are cluster wise differences in the mean values of CWQI, which means the spatial variability and clustering of bore wells based on physio-chemical parameters of the bore wells perform a very crucial role in the study area.

The ANOVA test findings are presented in Table 5. The test results evince that F-statistic ($F = 134.55$) and p-value ($p = 0.000$) is less than $\alpha = 0.05$, implying that there are critical differences in the average values of CWQI among the clusters. Further the difference in the mean values of CWQI, results in the for-

mation of clusters and are mainly due to anthropogenic activities viz. 1) Contamination based on indiscriminate MSW dumping and 2) Partially treated or SWW land application.

4.4. Multi-Linear Regression Model (MLR)

Regression models are best fit for establishing association between dependent and independent variables of small sample size. The MLR is a method applied to ascertain relationship between a dependent variable and one or more independent variables in a linear fashion and it is based on the method of least squares [38] [39]. In the best model, sum of the squared error between observed and predicted values of the parameters should be minimum. CWQI estimation also can be performed using MLR models which explain linear relationship among various hydrogeological parameters and is as follows:

$$Y = \alpha + a(\mathrm{EC}) + b(\mathrm{TDS}) + c(\mathrm{TH}) + d\left(\mathrm{HCO_3^-}\right) + e\left(\mathrm{Cl^-}\right) + f\left(\mathrm{SO_4^{2-}}\right) + g\left(\mathrm{Na^+}\right) + h\left(\mathrm{Ca^{2+}}\right) + i\left(\mathrm{Mg^{2+}}\right) + j\left(\mathrm{K^+}\right) + €$$

where, Y—CWQI ; α—regression constant ; €—random error.

a, b, c, d, e, f, g, h, i and j are coefficients of predictors in linear regression model;

EC, TDS, TH, $\mathrm{HCO_3^-}$, $\mathrm{Cl^-}$, $\mathrm{SO_4^{2-}}$, $\mathrm{Na^+}$, $\mathrm{Ca^{2+}}$, $\mathrm{Mg^{2+}}$ and $\mathrm{K^+}$ are input parameters.

The findings of MLR study for three clusters adopting 10 independent water quality parameters are summarized in Table 6. These are unstandardized regression co-efficients/weights which are incorporated in the regression equation. The MLR model developed employing stepwise regression technique to predict CWQI in Cluster 1 is:

$$CWQI - C1 = 122.39 - 0.006(\mathrm{EC}) - 0.016(\mathrm{TDS}) - 0.148(\mathrm{TH}) - 0.031\left(\mathrm{HCO_3^-}\right) - 0.077\left(\mathrm{Cl^-}\right) + 0.232\left(\mathrm{SO_4^{2-}}\right) - 0.063\left(\mathrm{Na^+}\right) + 0.18\left(\mathrm{Ca^{2+}}\right) + 0.037\left(\mathrm{Mg^{2+}}\right) + 0.37\left(\mathrm{K^+}\right)$$

Similarly MLR models for Clusters 2 and 3 can be developed. In Cluster 2, the variables EC, TDS and $\mathrm{Ca^{2+}}$ had been removed from the model during stepwise regression because their regression co-efficients were observed to be statistically inconsequential in simulating CWQI. Further F-test was adopted to examine the complete significance of the formulated MLR model for simulating CWQI. Furthermore the results of ANOVA of the MLR models of the three clusters are presented in Table 5.

From Table 5 & Table 6, it is observed that in Cluster 1 from the F-statistic (F= 12.165) and p value (p = 0.000), it is resolved that it is in fact a significant model i.e. the independent variables interpret a significant degree of variability in the prediction of CWQI and it is also confirmed that R^2 (0.88) is remarkably significant for this model.

It is evident from the Table 6 that the MLR model for Cluster 2 has the highest R^2 (1) value and is supported by F-statistic and p-level. However the MLR

Table 5. ANOVA test for cluster classification and MLR models.

		Sum of Squares	df	Mean Square	F	Sig.
Clusters CWQI	Between Groups	26723.167	2	13361.583	134.548	0.000
	Within Groups	6454.951	65	99.307		
Cluster 1 MLR	Between Groups	4300.274	10	430.027	12.165	0.000
	Within Groups	600.932	17	35.349		
Cluster 2 MLR	Between Groups	85.555	7	12.222		
	Within Groups	0.000	0			
Cluster 3 MLR	Between Groups	823.823	10	82.382	2.843	0.021
	Within Groups	608.584	21	28.980		

Table 6. MLR model co-efficients for selected water quality parameters.

Dependent Variable "Y"	Constant "α"	Independent Variable										R	R^2
		EC "a"	TDS "b"	TH "c"	HCO_3^- "d"	Cl^- "e"	SO_4^{2-} "f"	Na^+ "g"	Ca^{2+} "h"	Mg^{2+} "I"	K^+ "j"		
CWQI-C1	122.394	−0.006	−0.016	−0.148	−0.031	−0.077	0.232	−0.063	0.180	0.037	0.370	0.94	0.88
CWQI-C2	24.718	--	--	−0.131	0.034	0.067	0.016	0.016	--	0.187	−0.057	1	1
CWQI-C3	98.920	0.044	−0.033	0.056	−0.059	−0.283	−0.074	0.342	−0.240	−0.185	−3.508	0.76	0.58

model for Cluster 3 has the lowest R^2 (0.58) value. This indicates that CWQI of Cluster 3 is not influenced by the anthropogenic activities under consideration. Thus the values of R^2, F-statistic, and p-level for MLR models for all the three clusters are statistically significant which indicate that the formulated MLR models can simulate/predict CWQI reasonably.

In general the "sig" column in Table 5 provides the computed value of "p", if $p < 0.05$ then null hypothesis H_o is rejected and alternate hypothesis H_1 is accepted. In other words the independent variables are significant. In case of CWQI, it can be interpreted that there is spatial variability in Clusters 1, 2 and 3. Similarly the MLR models for Clusters 1, 2 and 3 are significant as the p values are <0.05, and so the alternate hypothesis H_1 is accepted and H_o is rejected.

4.5. Development of ANN-MLP Models

To simulate CWQI of groundwater, 10 most significant physio-chemical parameters were chosen. Multi Layer Perceptron (MLP) methodology of ANN was applied using SPSS Version 21.0. 1065 groundwater samples were analysed to model CWQI. 70% of the samples were utilized to train the ANN models and the balance 30% of data were employed to evaluate the model. Primarily 10 variables were used as inputs to ANN. To select the best fit ANN model, a methodology has been worked out for periodic removal of input parameters. By eliminating the input parameters the structure of optimized ANN model was made to run again.

4.5.1. Sensitivity Analysis

Sensitivity analyses were carried out for the ANN model to ascertain the relative weight of each input variable for reasonably simulating CWQI. This analysis was employed for all the three clusters by making certain modifications on distinct inputs and examining their consequences on the model output. The modification in the input was designed by removing certain parameters, while keeping the other input parameters intact and then the model output was simulated. Pursuing this removal approach 9 ANN models for Cluster 1 were developed and furnished in Table 7.

The same approach was adopted for Clusters 2 and 3. The optimal simulated results of all the 9 ANN and MLR models in all the three clusters are summarized in Tables 8-10.

4.5.2. Comparative Performance of the ANN and MLR Models

Further the performances of nine ANN models in simulating CWQI were collated with those of the corresponding MLR models by using R^2, RMSE and MAE (quantitative indicators) [40] [41] [42] [43]. The results of this comparison are also presented in Tables 8-10. The model which indicated high R^2 value and considerably low MAE and RMSE values, was considered to be best simulated model and is suitable for further analysis.

From the Tables 8-10 it could be seen that in Cluster 1, five ANN models showed R^2 values more than 0.9 and the remaining 4 models showed R^2 values between 0.85 and 0.9. The MLR model showed R^2 value as 0.88. Based on the RMSE (3.99) and MAE (3.52) values, the ANN 9 and MLR models could be considered for further analysis.

In Cluster 2, the R^2 values ranged from 0.175 to 0.995. The ANN 2 and ANN 6 models showed R^2 values as low as 0.175 and 0.232 respectively. Interestingly the MLR and ANN 8 models showed exactly the same values of R^2 (0.995), RMSE (0.23) and MAE (0.207). Both of these models could be considered for further investigation.

Table 7. Combination of input parameters in ANN models (Cluster 1).

Sl. No.	Model	ANN Architecture	No. of parameters	Combination of input parameters	Output parameter
1	ANN1	10-3-1	10	EC, TDS, TH, HCO_3^-, Cl^-, SO_4^{2-}, Na^+, Ca^{2+}, Mg^{2+} and K^+	CWQI
2	ANN2	8-2-1	8	EC, TDS, TH, HCO_3^-, Cl^-, SO_4^{2-}, Na^+ and Ca^{2+}	CWQI
3	ANN3	7-2-1	7	EC, TDS, TH, HCO_3^-, Cl^-, SO_4^{2-} and Na^+	CWQI
4	ANN4	6-2-1	6	EC, TDS, TH, HCO_3^-, Cl^- and Na^+	CWQI
5	ANN5	5-3-1	5	EC, TDS, TH, Cl^- and Na^+	CWQI
6	ANN6	4-4-1	4	EC, TDS, Cl^- and Na^+	CWQI
7	ANN7	6-3-1	6	EC, TDS, TH, HCO_3^-, Cl^-, and SO_4^{2-}	CWQI
8	ANN8	6-1-1	6	EC, TDS, Na^+, Ca^{2+}, Mg^{2+} and K^+	CWQI
9	ANN9	8-2-1	8	TH, HCO_3^-, Cl^-, SO_4^{2-}, Na^+, Ca^{2+}, Mg^{2+} and K^+	CWQI

Table 8. MLR and ANN models—Cluster 1 (CWQI).

BW	OBSERVED	MLR	ANN1	ANN2	ANN3	ANN4	ANN5	ANN6	ANN7	ANN8	ANN9
BW1	32	40	40	40	42	37	38	37	39	40	38
BW5	40	41	42	39	44	40	42	42	40	40	36
BW14	39	38	41	39	37	34	41	49	41	39	39
BW15	41	40	46	38	39	40	44	45	46	39	44
BW40	39	40	37	39	38	37	38	36	36	39	36
BW41	38	30	36	37	38	35	37	36	38	38	35
BW42	31	35	36	37	37	35	38	38	36	40	37
BW43	55	49	53	47	47	44	46	48	46	53	53
BW47	73	65	68	67	70	71	66	68	67	69	69
BW48	48	50	42	48	47	46	48	45	47	42	44
BW49	48	54	52	54	51	49	51	48	50	49	51
BW51	40	39	37	38	38	37	39	38	39	39	37
BW52	47	44	39	43	43	43	45	46	45	41	39
BW53	48	53	51	52	53	51	50	48	50	51	55
BW54	49	51	43	49	49	49	49	47	48	47	48
BW57	64	61	63	65	68	64	59	57	64	61	62
BW58	49	55	48	56	56	54	54	50	55	50	51
BW59	72	67	67	69	68	67	63	68	70	69	68
BW60	66	69	67	69	69	72	67	72	69	69	70
BW62	65	64	65	67	66	65	62	57	64	66	67
BW63	74	67	69	68	68	73	64	71	71	67	69
BW67	38	33	37	37	40	36	37	35	36	38	35
BW68	39	40	37	38	45	38	39	38	42	39	37
BW69	65	63	64	64	64	64	62	63	61	68	66
BW70	57	61	54	62	64	62	59	58	61	59	59
BW71	65	64	68	66	64	66	64	71	65	71	68
BW72	73	73	66	70	72	73	68	74	71	72	70
BW78	49	59	53	62	65	62	60	61	59	58	56
R^2		0.88	0.91	0.88	0.85	0.90	0.88	0.86	0.91	0.91	0.91
RMSE		4.64	4.15	4.65	5.22	4.41	4.89	4.95	4.07	4.08	3.99
MAE		3.74	3.51	3.48	3.93	3.14	3.74	3.95	3.10	3.02	3.52

Note: BW—Borewell. All values in the Table indicate observed and simulated values of CWQI.

Table 9. MLR and ANN models—Cluster 2 (CWQI).

BW	OBSERVED	MLR	ANN1	ANN2	ANN3	ANN4	ANN5	ANN6	ANN7	ANN8	ANN9
BW7	36	36	34	29	35	31	35	29	36	36	32
BW8	28	28	29	28	28	28	29	29	28	28	28
BW9	28	28	28	28	28	28	28	29	30	28	29
BW10	28	28	28	28	28	28	28	29	28	28	28
BW11	35	35	35	29	34	34	35	28	35	35	35
BW13	29	29	30	29	29	29	29	29	29	29	29
BW44	29	29	29	29	29	27	29	29	28	29	27
BW45	34	34	34	28	32	34	34	29	31	34	32
R^2		0.995	0.945	0.175	0.970	0.743	0.985	0.232	0.823	0.995	0.778
RMSE		0.230	0.885	3.949	0.968	1.986	0.500	4.024	1.418	0.230	1.841
MAE		0.207	0.653	2.563	0.688	1.188	0.403	2.864	0.917	0.207	1.273

Note: BW—Borewell. All values in the Table indicate observed and simulated values of CWQI.

Table 10. MLR and ANN models—Cluster 3 (CWQI).

BW	OBSERVED	MLR	ANN1	ANN2	ANN3	ANN4	ANN5	ANN6	ANN7	ANN8	ANN9
BW17	80	80	85	81	87	84	86	85	83	84	87
BW18	89	87	87	88	89	88	85	87	88	86	88
BW19	58	67	80	61	59	82	71	83	82	82	67
BW20	83	82	82	83	81	85	84	85	86	83	82
BW21	83	86	86	87	83	87	86	87	88	85	88
BW22	86	83	87	86	85	87	85	87	89	85	88
BW23	86	90	87	87	85	87	85	87	88	85	88
BW24	83	86	86	84	83	87	85	86	87	85	88
BW25	85	87	87	86	84	87	86	87	88	85	88
BW26	84	87	86	85	83	86	86	86	86	85	88
BW27	92	82	86	86	89	86	85	86	87	85	87
BW28	92	84	84	85	91	86	84	86	87	84	87
BW30	86	88	87	87	84	87	86	87	88	85	88
BW32	83	86	80	87	84	84	81	83	82	83	84
BW35	83	81	86	84	83	87	85	87	88	85	88
BW36	92	87	85	88	89	87	86	86	88	85	87
BW37	84	88	86	87	83	87	86	87	88	85	88
BW38	83	83	81	84	81	83	85	83	83	83	83
BW39	90	94	87	91	90	88	87	87	89	86	88
BW46	83	79	81	86	84	83	85	83	83	83	83
BW50	91	82	83	88	93	81	86	83	82	85	87
BW55	72	76	81	74	71	81	74	81	82	84	72
BW64	87	88	87	87	87	87	86	87	87	86	88
BW75	92	91	81	86	94	85	75	83	82	83	88
BW76	92	86	82	88	92	83	85	83	82	84	87
BW77	74	77	80	72	72	81	69	81	82	83	68
BW79	89	87	87	88	89	88	85	87	89	86	88
BW80	84	91	86	88	84	86	87	86	85	85	88
BW81	82	84	85	86	83	85	86	86	82	85	88
BW82	86	88	87	87	85	87	86	87	86	85	88
BW84	92	86	84	91	88	83	87	84	82	85	87
BW85	83	86	86	87	83	85	87	86	84	85	88
R^2		0.555	0.178	0.806	0.925	0.193	0.414	0.148	0.122	0.282	0.646
RMSE		4.463	6.074	2.967	1.867	6.104	5.190	6.248	6.326	6.206	4.057
MAE		3.678	4.341	2.347	1.309	4.091	3.834	4.247	4.159	4.053	3.441

Note: BW—Borewell. All values in the Table indicate observed and simulated values of CWQI.

In Cluster 3, the R^2 values ranged from 0.122 to 0.925. The ANN 1 and ANN 4 to ANN 8 models showed low R^2 values. The ANN 3 is the only model which showed R^2 value as high as 0.925 while MLR showed R^2 value as 0.555. So ANN 3

model could be termed as best fit model for further examination.

Conclusively in the anthropogenically polluted Clusters 1 and 2, both MLR and ANN models could be considered for further investigation. But, in Cluster 3 which is less polluted, only ANN model is best fit for simulation.

The graphical comparison of observed and optimal simulated CWQI by ANN and MLR models in all the 3 clusters are depicted in Figures 3-5. It is evident from these figures that the predicted CWQI derived by both MLR and ANN models tally fairly well with the observed CWQI in Clusters 1 and 2, whereas in Cluster 3 only ANN model fits well with the observed values. In addition to the concurrent plots, the comparison between observed and simulated CWQIs by ANN was analysed by scatter plots with 1:1 equilines and error bands for all the 3 clusters and the same are illustrated in Figures 6-8.

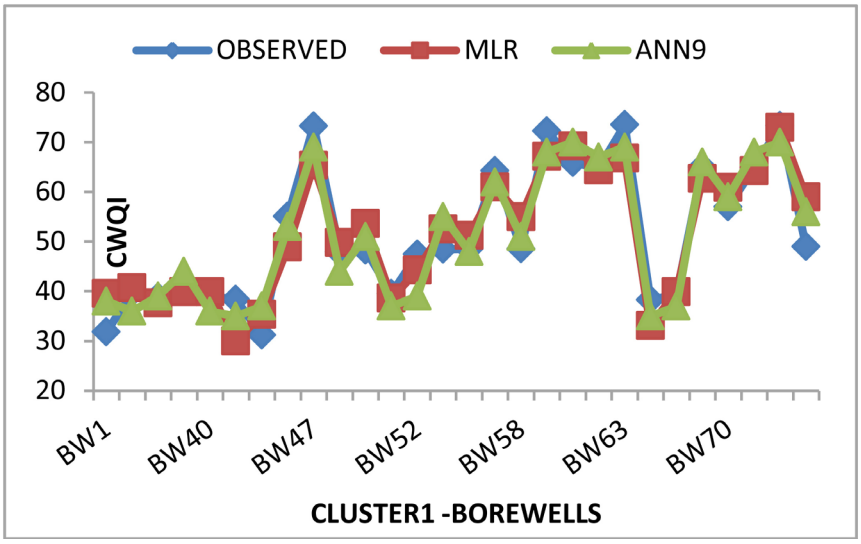

Figure 3. Observed and predicted CWQI in Cluster 1.

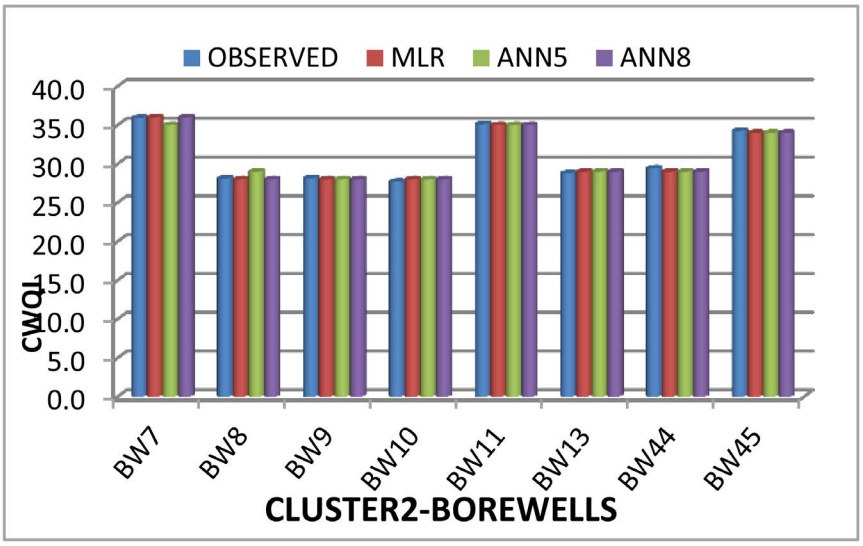

Figure 4. Observed and predicted CWQI in Cluster 2.

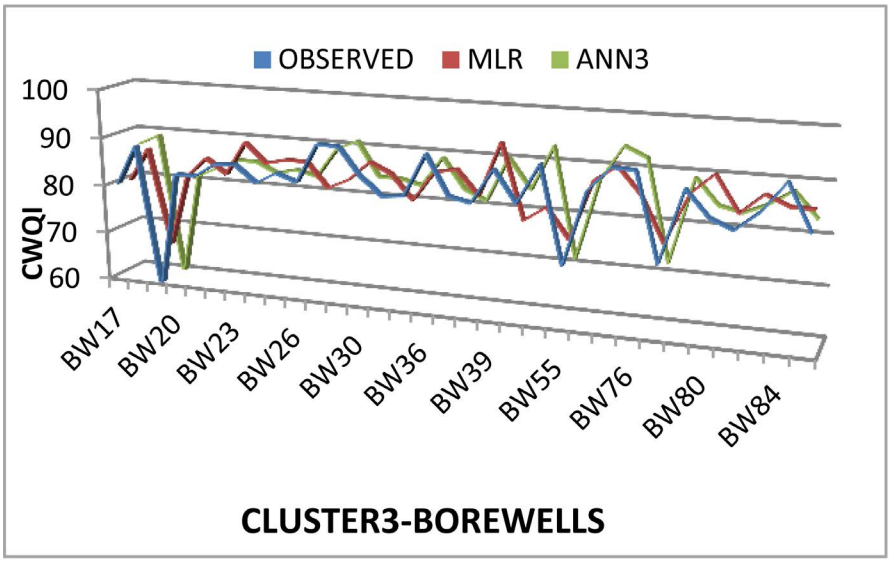

Figure 5. Observed and predicted CWQI in Cluster 3.

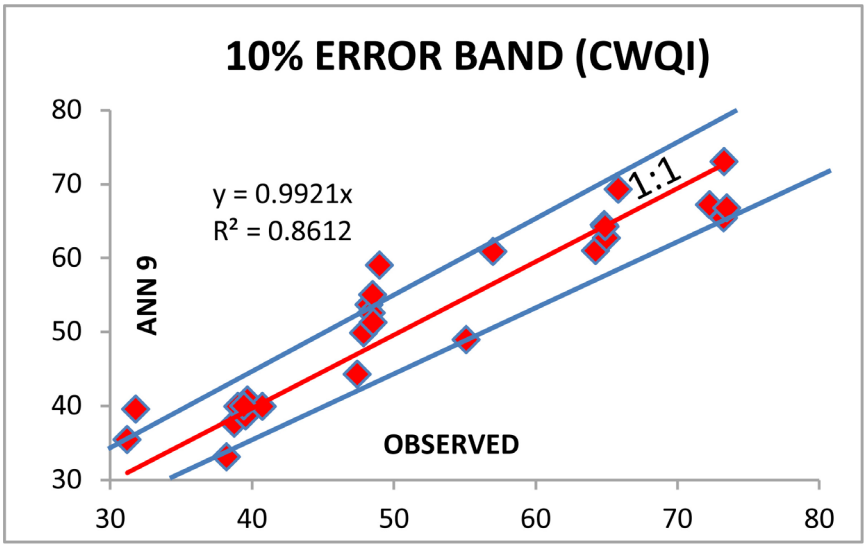

Figure 6. Error band in Cluster 1.

In these figures the parallel lines (qualitative indicators) indicate higher and lower error bands in relation to 1:1 line. Understandably the simulated CWQI of 23 borewells (82%) fall within ±10% error band in Cluster 1. In Clusters 2and 3, 100% and 94% of the borewells fall within ±5% error band respectively. In view of the quantitative and qualitative realization gauges, the ANN models outperform MLR models in a much better way and this could be ascribed to the fact that MLR is dependent on method of least squares and it is linear in nature, whereas ANN is based on sophisticated nonlinear methods.

5. Conclusion

This research work was attempted to investigate the strength of two data induced methodologies *viz*. MLR and ANN, for predicting CWQI in groundwater

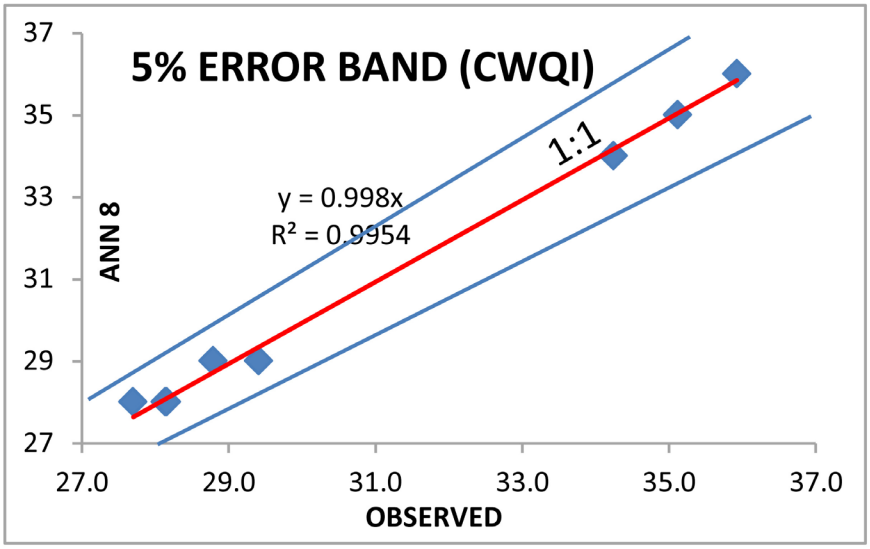

Figure 7. Error band in Cluster 2.

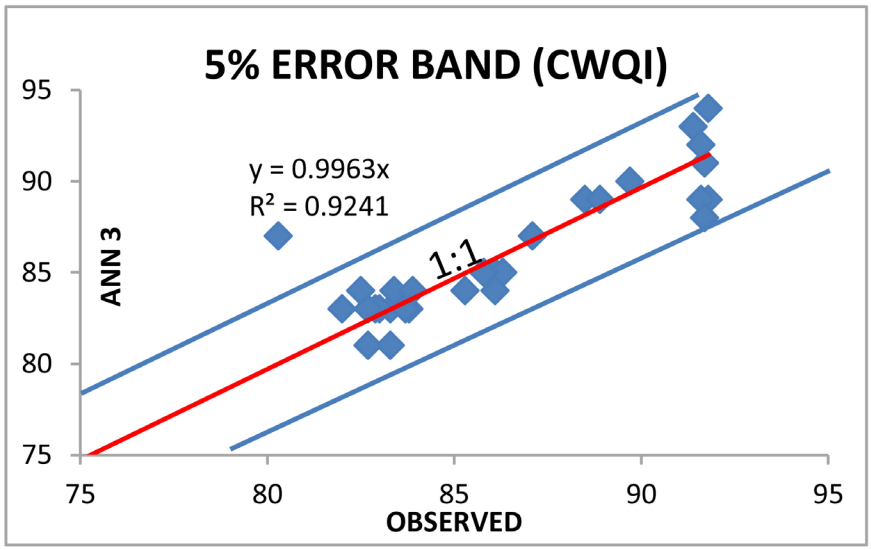

Figure 8. Error band in Cluster 3.

quality scenario. Using HCA, 3 clusters were developed based on 10 most significant physio-chemical parameters. The cluster classification based on anthropogenic activities and MLR models so developed have been validated by ANOVA using F-statistic. Each cluster was subjected to simulation of CWQI using one MLR and nine ANN models. The quantitative and qualitative performances of MLR and ANN models were assessed statistically and graphically. The analyses of the results revealed that both MLR and ANN models were fairly good in predicting CWQI in Clusters 1 and 2 with high R^2, low RMSE and MAE values. In Cluster 3 only ANN model fared well. This is because MLR models are generally linear in nature and they are poor in their prediction ability especially in noisy data environment. Therefore, ANN models can be considered as a powerful and dependable tool in simulating complex, inter-dependable physio-chemical pa-

rameters. Based on the required information, it will be very simple, clear and convenient to adopt either MLR or ANN model depending on environmental conditions to predict the water quality of the study area in a more practical way so that the associated costs, sampling points etc., can be minimized to the maximum extent possible. Thus this research attempt will be very practical to the groundwater management community which is involved in water quality problems.

References

[1] Tiwari, A.K. and Singh, A.K. (2014) Hydrogeochemical Investigation and Groundwater Quality Assessment of Pratapgarh District, Uttar Pradesh. *Journal of the Geological Society of India*, **83**, 329-343. https://doi.org/10.1007/s12594-014-0045-y

[2] Tiwari, A.K., De Maio, M., Singh, P.K. and Singh, A.K. (2016) Hydrogeochemical Characterization and Groundwater Quality Assessment in a Coal Mining Area, India. *Arabian Journal of Geosciences*, **9**, 177. https://doi.org/10.1007/s12517-015-2209-5

[3] Chandra, S., Singh, P.K., Tiwari, A.K., Panigrahy, B.P. and Kumar, A. (2015) Evaluation of Hydrogeological Factor and Their Relationship with Seasonal Water Table Fluctuation in Dhanbad District, Jharkhand, India. *ISH Journal of Hydraulic Engineering*, **21**, 193-206. https://doi.org/10.1080/09715010.2014.1002542

[4] Tiwari, A.K., Singh, P.K., Singh, A.K. and De Maio, M. (2016) Estimation of Heavy Metal Contamination in Groundwater and Development of a Heavy Metal Pollution Index by Using GIS Technique. *Bulletin of Environmental Contamination and Toxicology*, **96**, 508-515. https://doi.org/10.1007/s00128-016-1750-6

[5] Hildenbrand, Z.L., *et al*. (2015) A Comprehensive Analysis of Ground Water Quality in the Barenett Shale Region. *Environmental Science & Technology*, **49**, 8254-8262. https://doi.org/10.1021/acs.est.5b01526

[6] Peters, N.E. and Michel, M. (2000) Water Quality Degradation Effects on Fresh Water Availability. *Impacts to Human Activities*, **25**, 185-193.

[7] Tiwari, A.K., Singh, A.K., Singh, A.K. and Singh, M.P. (2015) Hydro Geochemical Analysis and Evaluation of Surface Water Quality of Pratapgarh District, Uttar Pradesh, India. *Applied Water Science*, 1-15.

[8] Suresh, N., Saravanane, R. and Sundararajan, T. (2016) Assessment of Groundwater Contamination Due to Co-Disposal of Municipal Solid Waste and Secondary Wastewater on Land at Lawspet in Puducherry, India. *International Journal of Environmental Engineering and Management*, **7**, 35-67.

[9] APHA (1998) Standard Methods for the Examination of Water and Wastewater. 20th Edition, American Public Health Association, Washington DC.

[10] Sharma, J.D., *et al*. (2005) Chemical Analysis of Drinking Water of Sanganer Tehsil, Jaipur District. *International Journal of Environmental Science and Technology*, **2**, 373-379.

[11] Sarwar, M.I., Majumderb, A.K. and Islam, M.N. (2010) Water Quality Parameters: A Case Study of Karnafully River, Chittagong, Bangladesh. *Bangladesh Journal of Scientific and Industrial Research*, **45**, 177-181. https://doi.org/10.3329/bjsir.v45i2.5722

[12] Rahman, L., Islam, M., Hossain, M.Z. and Ahsan, M.A. (2012) Study of the Seasonal Variations in Turag River Water Quality Parameters. *African Journal of Pure and Applied Chemistry*, **6**, 144-148.

[13] Danquah, L., Abass, K. and Nikoi, A.A. (2011) Antropogenic Pollution of Inland Waters: The Case of the Aboabo River in Kumasi, Ghana. *Journal of Sustainable Development*, **4**, 103. https://doi.org/10.5539/jsd.v4n6p103

[14] Amadi, A.N., *et al*. (2012) Geostatistical Assessment of Groundwater Quality from Coastal Aquifers of Eastern Niger Delta, Nigeria. *Geosciences*, **2**, 51-59.

[15] Batayneh, A. and Zumlot, T. (2012) Multivariate Statistical Approach to Geochemical Methods in Water Quality Factor Identification; Application to the Shallow Aquifer System of the Yarmouk Basin of North Jordan. *Research Journal of Environmental and Earth Sciences*, **4**, 756-768.

[16] Cobbina, S.J., *et al*. (2012) Multivariate Statistical and Spatial Assessment of Groundwater Quality in the Tolon-Kumbungu District, Ghana. *Research Journal of Environmental and Earth Sciences*, **4**, 88-98.

[17] Nosrati, K. and Van Den Eeckhaut, M. (2012) Assessment of Groundwater Quality using Multivariate Statistical Techniques in Hashtgerd Plain, Iran. *Journal of Environmental Earth Science*, **65**, 331-344. https://doi.org/10.1007/s12665-011-1092-y

[18] Singh, R., Vishal, V., Singh, T.N. and Ranjith, P.G. (2012) A Comparative Study of Generalized Regression Neural Network Approach and Adaptive Neuro-Fuzzy Inference Systems for Prediction of Unconfined Compressive Strength of Rocks. *Neural Computing and Applications*, **23**, 499-506. https://doi.org/10.1007/s00521-012-0944-z

[19] Singh, R., Vishal, V. and Singh, T.N. (2012) Soft Computing Method for Assessment of Compressional Wave Velocity. *Scientia Iranica—Transactions in Civil Engineering*, **19**, 1018-1024.

[20] Affandi, A.K., Watanabe, K. and Tirtomihardjo, H. (2007) Application of an Artificial Neural Network to Estimate Groundwater Level Fluctuation. *Journal of Spatial Hydrology*, **7**, 23-46.

[21] Feng, S., Kang, S., Huo, Z., Chen, S. and Mao, X. (2008) Neural Networks to Simulate Regional Ground Water Levels Affected by Human Activities. *Ground Water*, **46**, 80-90.

[22] Gerken, W.C., Purvis, L.K. and Butera, R.J. (2006) Genetic Algorithm for Optimization and Specification of a Neuron Model. *Neurocomputing*, **69**, 1039-1042. https://doi.org/10.1016/j.neucom.2005.12.041

[23] Hani, A., Lallahem, S., Mania, J. and Djabri, L. (2006) On the Use of Finite Difference and Neural-Network Models to Evaluate the Impact of Underground Water Overexploitation. *Hydrological Processes*, **20**, 4381-4390. https://doi.org/10.1002/hyp.6173

[24] Krishna, B., Rao, Y.R.S. and Vijaya, T. (2008) Modeling Groundwater Levels in an Urban Coastal Aquifer Using Artificial Neural Networks. *Hydrological Processes*, **22**, 1180-1188. https://doi.org/10.1002/hyp.6686

[25] Lallahem, S., Mania, J., Hani, A. and Najjar, Y. (2005) On the Use of Neural Networks to Evaluate Groundwater Levels in Fractured Media. *Journal of Hydrology*, **307**, 92-111. https://doi.org/10.1016/j.jhydrol.2004.10.005

[26] Sarala, T.D. and Uma Maheswari, T.S.R. (2014) Modeling of Irrigation Water Quality Using Multilayer Perceptron Back Propagation Neural Network (MLBP-NN). *International Journal of Chem Tech Research*, **6**, 3053-3061.

[27] Mohanty, S., Jha, M.K., Kumar, A. and Sudheer, K.P. (2009) Artificial Neural Network Modeling for Groundwater Level Forecasting in a River Island of Eastern India. *Water Resources Management*, **24**, 1845-1865. https://doi.org/10.1007/s11269-009-9527-x

[28] Nayak, P.C., Rao, Y.R.S. and Sudheer, K.P. (2006) Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. *Water Resources Management*, **20**, 77-90. https://doi.org/10.1007/s11269-006-4007-z

[29] Nikolos, I.K., Stergiadi, M., Papadopoulou, M.P. and Karatzas, G.P. (2008) Artificial Neural Networks as an Alternative Approach to Groundwater Numerical Modeling and Environmental Design. *Hydrological Processes*, **22**, 3337-3348. https://doi.org/10.1002/hyp.6916

[30] Nourani, V., Mogaddam, A.A. and Nadiri, A.O. (2008) An ANN-Based Model for Spatiotemporal Groundwater Level Forecasting. *Hydrological Processes*, **22**, 5054-5066. https://doi.org/10.1002/hyp.7129

[31] Sethi, R.R., Kumar, A., Sharma, S.P. and Verma, H.C. (2010) Prediction of Water Table Depth in a Hard Rock Basin by Using Artificial Neural Network. *International Journal of Water Resources and Environmental Engineering*, **2**, 95-102.

[32] Kamble, S.R. and Vijay, R. (2011) Assessment of Water Quality Using Cluster Analysis in Coastal Region of Mumbai, India. *Environmental Monitoring and Assessment*, **178**, 321-332. https://doi.org/10.1007/s10661-010-1692-0

[33] Shihab, A.S. and Hashim, A. (2006) Cluster Analysis Classification of Groundwater Quality in Wells within and around Mosul City, Iraq. *Journal of Environmental Hydrology*, **14**, 1-11.

[34] Lin, C., Wu, E.M.Y., Lee, C.N. and Kuo, S.L. (2010) Multivariate Statistical Factor and Cluster Analyses for Selecting Food Waste Optimal Recycling Methods. *Environmental Engineering Science*, **28**, 349-356. https://doi.org/10.1089/ees.2010.0158

[35] Devi Prasad, A.G. and Kothathi, S. (2012) Application of CCME Water Quality Index to the Lakes of Mandya, Karnataka State, India. *Online International Interdisciplinary Research Journal*, **2**, 1.

[36] Kuruppu, U. and Rahman, A. (2015) Trends in Water Quality Data in the Hawkesbury-Nepean River System, Australia. *Journal of Water and Climate Change*, **10**, 2166.

[37] Canadian Council of Ministers of the Environment (2001) Canadian Water Quality Guidelines for the Protection of Aquatic Life. Canadian Council of Ministers of the Environment, Canada.

[38] Chenini, I. and Khemiri, S. (2009) Evaluation of Groundwater Quality Using Multiple Linear Regressions and Structural Equation Modeling. *International Journal of Environmental Science and Technology*, **6**, 509-519. https://doi.org/10.1007/BF03326090

[39] Sinnakaudan, S.K., Ghani, A.A., Ahmad, M.S.S. and Zakaria, N.A. (2006) Multiple Linear Regression Model for Total Bed Material Load Prediction. *Journal of Hydraulic Engineering*, **132**, 521-528. https://doi.org/10.1061/(ASCE)0733-9429(2006)132:5(521)

[40] Adeloye, A.J. (2009) Multiple Linear Regression and Artificial Neural Network Models for Generalized Reservoir Storage-Yield-Reliability Function for Reservoir Planning. *Journal of Hydraulic Engineering*, **14**, 731-738. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000041

[41] Heuvelmans, G., Muys, B. and Feyen, J. (2006) Regionalisation of the Parameters of a Hydrological Model: Comparison of Linear Regression Models with Artificial Neural Nets. *Journal of Hydrology*, **319**, 245-265. https://doi.org/10.1016/j.jhydrol.2005.07.030

[42] Sarala Thambavani, D. and UmaMaheswari, T.S.R. (2015) Application of Multivariate Linear Regression and Neural Network in the Assessment of Ground Water

Quality. *International Journal of Chem Tech Research*, **8**, 1282-1289.

[43] Sahoo, S. and Jha, M.K. (2013) Groundwater-Level Prediction Using Multiple Linear Regression and Artificial Neural Network Techniques: A Comparative Assessment. *Hydrogeology Journal*, **21**, 1865-1887. https://doi.org/10.1007/s10040-013-1029-5