

Retraction Notice

Title of retracted article: **Cluster Analysis to Assess Groundwater Quality in Erode District, Tamil Nadu, India**

Author(s): Ganeshbabu Oorkavalan, Sashikkumar Madurai Chidambaram, Vijayaraj Mariappan, Gokulakrishnan Kandaswamy, Sakthieswaran Natarajan

* Corresponding author. Email: oganeshbabu@gmail.com

Journal: Circuits and Systems
 Year: 2016
 Volume: 7
 Number: 6
 Pages (from - to): 877 - 890
 DOI (to PDF): <http://dx.doi.org/10.4236/cs.2016.76075>
 Paper ID at SCIRP: 66545
 Article page: <http://www.scirp.org/Journal/PaperInformation.aspx?PaperID=66545>
 Retraction date: 2017-03-24

Retraction initiative (multiple responses allowed; mark with X):

- All authors
- Some of the authors:
- Editor with hints from
- Journal owner (publisher)
- Institution:
- Reader:
- Other:
- Date initiative is launched: 2017-03-16

Retraction type (multiple responses allowed):

- Unreliable findings
- Lab error
- Inconsistent data
- Analytical error
- Biased interpretation
- Other:
- Irreproducible results
- Failure to disclose a major competing interest likely to influence interpretations or recommendations
- Unethical research
- Fraud
- Data fabrication
- Fake publication
- Other:
- Plagiarism
- Self plagiarism
- Overlap
- Redundant publication *
- Copyright infringement
- Other legal concern:
- Editorial reasons
- Handling error
- Unreliable review(s)
- Decision error
- Other:
- Other:

Results of publication (only one response allowed):

- are still valid.
- were found to be overall invalid.

Author's conduct (only one response allowed):

- honest error
- academic misconduct
- none (not applicable in this case – e.g. in case of editorial reasons)

* Also called duplicate or repetitive publication. Definition: "Publishing or attempting to publish substantially the same work more than once."

History

Expression of Concern:

yes, date: yyyy-mm-dd

no

Correction:

yes, date: yyyy-mm-dd

no

Comment:

The paper does not meet the standards of "Circuits and Systems".

This article has been retracted to straighten the academic record. In making this decision the Editorial Board follows [COPE's Retraction Guidelines](#). Aim is to promote the circulation of scientific research by offering an ideal research publication platform with due consideration of internationally accepted standards on publication ethics. The Editorial Board would like to extend its sincere apologies for any inconvenience this retraction may have caused.

Cluster Analysis to Assess Groundwater Quality in Erode District, Tamil Nadu, India

Ganeshbabu Oorkavalan¹, Sashikkumar Madurai Chidambaram², Vijayaraj Mariappan³, Gokulakrishnan Kandaswamy⁴, Sakthieswaran Natarajan¹

¹Department of Civil Engineering, Anna University Regional Campus, Tirunelveli, India

²Department of Civil Engineering, University College of Engineering, Dindigul, India

³Department of Electronics and Communication Engineering, GCE, Tirunelveli, India

⁴Department of E & C Engineering, Anna University Regional Campus, Tirunelveli, India

Email: oganeshbabu@gmail.com

Received 31 March 2016; accepted 13 May 2016; published 18 May 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Water is a complicated environment system; traditional methods cannot meet the demands of water environment protection. As the frontrunner of complex nonlinear science and artificial intelligence, artificial neural network has begun to be applied in the field of water quality evaluation and estimation. In view of the deficiency of the traditional methods, artificial intelligence techniques, such as neural networks modeling tools, can be applied to assess water quality parameters. This study is conducted to evaluate factors regulating groundwater quality in and around Erode District, Tamil Nadu, India. This investigation is focused on the determination of physico-chemical parameters such as pH, EC, TDS, Ca, Mg, TH, Na, K, HCO₃, SO₄ and Cl. Groundwater suitability for drinking, domestic and agricultural purposes is examined with WHO standards. Dominant factors controlling the hydro-geochemistry of groundwater in the study area is indicated by Principal Component Analysis. Classification methods are used to classify the water quality regulating factors. Cluster analysis is supporting for the grouping on the basis of contamination characteristics of groundwater quality. This study also reveals that multivariate statistical analyses are used to improve the understanding of groundwater condition and appraisal of groundwater quality.

Keywords

Groundwater, Water Quality, Principal Component Analysis, Classification, Multilayer Perceptron, Dendrogram

1. Introduction

Water is the nature's most essential wonderful, abundant and useful compound to the existence of human and all other living things and is used by human for domestic purposes including drinking, agriculture irrigation, and manufacturing industries. Quality of water is vital for the survival of humans and ecosystems and is the indication of wide variety of chemical and biological constituents present and also it is beneficial when these constituents are present within desirable levels. Groundwater is an important source of water supply throughout the world. Groundwater quality estimation is a part of environment assessment and is closely related with human wellbeing. The water quality is usually evaluated by measuring a broad range of parameters such as pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS), Total Hardness (TH), Bicarbonate (HCO_3), Calcium (Ca), Magnesium (Mg), Chloride (Cl), Sodium (Na), Fluoride (F), Sulphate (SO_4) and Potassium (K), and the concentrations of a variety of other pollutants such as pathogens, nutrients, organics and metals [1].

Water pollution is inevitably caused by swift economic growth. Automatic water quality monitoring system and accurate water quality forecast methods are very important to effectively regulate water pollution. Water quality assessment is an important monitoring project. Presently there are many methods existing to assess the water quality, such as Genetic Algorithm, Mathematical Statistics Method, Model-Based Approach, Gray Correlation Analysis method, Gray Clustering Method, Bayesian Approach and Principal Component Analysis [2]-[11]. Until now, many methods are used to predict water quality including Fuzzy Comprehensive Evaluation, Unascertained Measure, Integrated Pollution Index Method and Neural Network Model [12] [13]. Artificial Neural Networks (ANNs) are self-organizing, self-teaching, nonlinear methods and can deal with the systems which are difficult to be described with traditional methods [14]-[16]. ANNs are well-suited to predict water quality because of their informative processing characteristics, such as nonlinearity, parallelism, noise tolerance, learning and generalization capabilities, distributed association, fault-tolerance ability and applicability to complex problem [17]-[19].

Novel computational approaches, such as artificial neural networks, have found wide acceptance in many disciplines and it provides an alternative method for understanding and managing the water quality in reservoirs. In recent years, many researchers have been conducted on the water quality assessment. Approximately, 85%-90% of the water quality prediction works have been completed using Neural Network. In natural environment, water quality is a multivariate phenomenon, at least as reflected in the multitude of constituents which are used to characterize the quality of water body. The ANNs are popular tools for modeling highly complicated relationships, processes and phenomena. Many factors affect water quality. These factors have complex non-linear relationship with water quality. The data dimensionality should be reduced to extract the most important factors. PCA is a technology that can compress multiple original indices into a few aggregative variable indices, which can represent original data information. PCA has been successfully applied in environmental data analysis [20] [21]. Here, PCA is applied to optimize and select the sample set.

However, the classification of physico-chemical parameters in the assessment of water quality using Multi-Layer Perceptron Classifier (MLP) with and without Principal Component Analysis using WEKA 3.7. Tool has not been reported. The main objective of this study is to compare MLP with and without PCA to classify water quality. By comparing the advantages of these algorithms, the water quality prediction system can not only ensure the prediction accuracy of water quality, but also can improve prediction speed. As the groundwater mining has been continuously increasing to keep pace with agricultural development in rural areas, understanding the potential manipulation by human activity on groundwater quality is the key for protection and sustainable use of groundwater resources. Hence the study was undertaken by randomly collected 56 groundwater samples from dug and deep wells before and after Pre-monsoon and Post-monsoon seasons in the years 2009 and 2011 in the months of January and July at different locations of Erode district and are analyzed for their physico-chemical characteristics to understand the sources of dissolved ions, and to assess the chemical quality of the groundwater through physico-chemical analysis.

Groundwater in the study area is utilized for both agricultural and drinking purposes. Many villages in the District of Erode are facing water quality problem as well as drinking water shortage, especially during pre-monsoon season. Adults and children of this region are suffering from health problems due to consumption of contaminated water. This paper assesses groundwater quality of Erode district region for determining its suitability for drinking purposes. In this paper, a new water quality assessment method is studied, which can be so much more effective and objective to overcome the shortcoming of the existing methods.

1.1. Description of the Study Area

The Erode district covers an area of approximately 5692 km². Erode district is landlocked and is situated at between 10°36" and 11°58" north latitude and between 76°49" and 77°58" east longitude as shown in **Figure 1**.

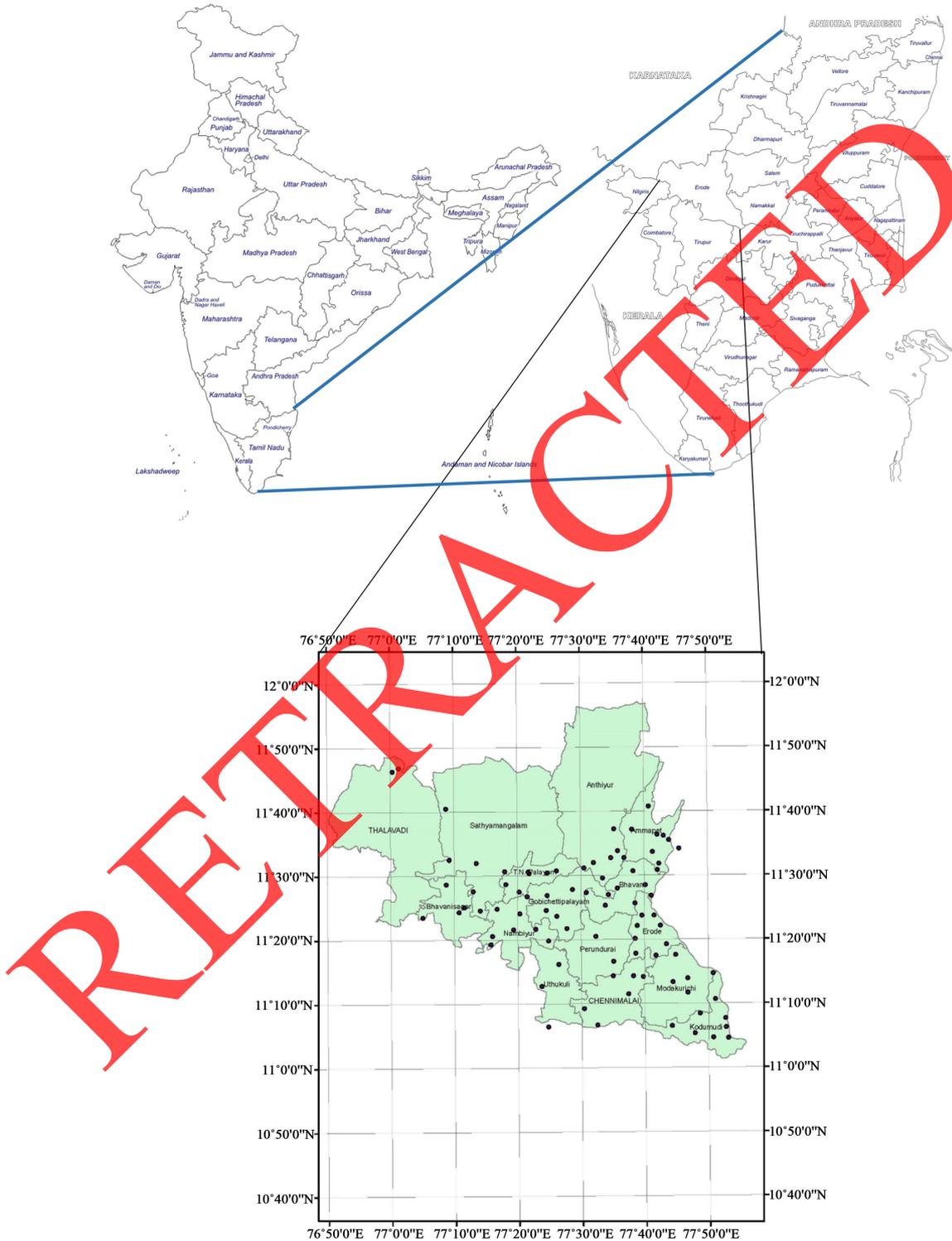


Figure 1. Location map of the study area—Erode district.

Erode district had a population of 2,251,744. The district had a total of 658,071 households. There are a total of 1,195,773 workers, comprising 173,376 cultivators, 331,414 main agricultural laborers, 48,960 in house hold industries, 557,301 other workers, 84,722 marginal workers, 4794 marginal cultivators, 38,798 marginal agricultural laborers, 5362 marginal workers in household industries and 35,768 other marginal workers as per 2011 census.

The district receives the rain under the influence of both southwest and northeast monsoons. The normal annual rainfall over the district varies from about 575 mm to 833 mm. The western part of the Erode district enjoys a salubrious climate because of the hilly region, whereas the central and eastern parts of the district are hot and humid. The relative humidity varies from 65 to 87 percent during the northeast monsoon period between October and November. Highest temperatures are recorded during the months of April and May.

The Erode district forms part of the uplands of the state. Physiographically the district can be divided into hilly area, the upland area and plains area. The prominent geomorphic units identified in the district through interpretation of Satellite imagery are 1) Structural hills, 2) Inselberg, 3) Ridges, 4) Valley fill, 5) Pediments, 6) Shallow Pediments. The soils of Erode district can be broadly classified into 6 major soils types viz., Red Calcareous Soil, Red Non Calcareous soil, Black Soil, Alluvial and Colluvial soils, Brown soil and Forest soil [22].

Erode district is underlain entirely by Archaean Crystalline formations with Recent alluvial deposits occurring along the river and streams courses and colluviums of valley-fills. The important aquifer systems in the district are constituted by weathered, fissured and fractured crystalline rocks and the recent alluvial deposits. The porous formations in the district are represented by alluvium and colluviums. The alluvial deposits are confined to the major river and stream courses only. Groundwater occurs under phreatic conditions. The maximum saturated thickness of these aquifers is up to 5 m depending upon the topographic conditions. The area lying at the foot hill zones which are seen in the northern parts of the district is underlain by the colluvial material derived from the nearby hill ranges comprising sands and gravels.

The maximum saturated thickness of these aquifers is up to 20 m depending upon the topographic conditions. Groundwater occurs under phreatic conditions. The hard consolidated crystalline rocks of Archaean age represent weathered, fissured and fractured formations of gneisses, granites, charnockites and other associated rocks. Groundwater occurs under phreatic conditions in the weathered mantle and under semi-confined conditions in the fractured zones.

1.2. Scope of the Study

The scope of the present study is to determine:

- The effect of physico-chemical parameters in the assessment of water quality during the years 2009 and 2011 using Correlation Analysis.
- Classification of physico-chemical parameters in the assessment of water quality using MLP classifier with and without Principal Component Analysis using WEKA 3.7.Tool.
- The order of influence of parameters or variables affecting the water quality using cluster analysis in Minitab 16 Statistical Software.

2. Method of Study

The study was designed to investigate the conditions of groundwater contamination in the study area. Samples were taken out in a pre-cleaned plastic polyethylene bottle. Prior to sampling, all the plastic bottles were washed and rinsed thoroughly with the groundwater. Water quality parameters such as pH and electrical conductivity (EC) were analyzed immediately. Other parameters such as Total Dissolved Solids (TDS), Total Hardness (TH), Bicarbonate (HCO_3), Calcium (Ca), Magnesium (Mg), Chloride (Cl), Sodium (Na), Fluoride (F), Sulphate (SO_4) and Potassium (K) were later analyzed in the Environmental Engineering Laboratories of Anna University Regional Campus, Tirunelveli.

3. Results and Discussion

The water quality investigation of different groundwater samples were carried out to determine different physico-chemical parameters such as for pH, EC, TDS, Ca, Mg, TH, Na, K, HCO_3 , SO_4 , and Cl; and the results are presented in **Table 1**. pH value of groundwater samples varied between 7.6 - 8.9 and 6.1 - 8.6 during the Pre-

Table 1. Physico-chemical parameters of groundwater in the study area.

Sl. No.	Water Quality Parameters	WHO Standard (1993)	Pre-Monsoon Season		Post-Monsoon Season	
			2009	2011	2009	2011
1	pH (mg/l)	6.5 - 8.5	7.6 - 8.9	7.7 - 8.9	6.1 - 8.6	8.0 - 9.0
2	Electrical Conductivity ($\mu\text{S}/\text{cm}$) (EC)	-	400 - 3190	345 - 3020	340 - 2920	260 - 4180
3	Total Dissolved Solids (TDS) (mg/l)	500 - 1500	221 - 1829	158 - 2093	209 - 2527	125 - 2515
4	Calcium (Ca) (mg/l)	200	6 - 160	10 - 168	8 - 136	8 - 152
5	Magnesium (Mg) (mg/l)	150	12.15 - 170.1	12.1 - 150.7	4 - 129	8.5 - 189.5
6	Sodium (Na) (ppm)	200	2 - 391	23 - 529	0 - 460	10 - 598
7	Potassium (K) (ppm)	-	5 - 219	2 - 145	0 - 192	4 - 375
8	Bicarbonate (HCO_3) (mg/l)	-	29.3 - 750.3	94.3 - 732	55 - 659	73.14 - 793
9	Sulphate (SO_4) (mg/l)	250	7 - 386	11 - 219	11 - 331	10 - 376
10	Chloride (Cl) (mg/l)	250 - 600	18 - 652	21 - 978	14 - 518	14 - 1106

monsoon and Post-monsoon seasons, respectively in the year 2009 whereas it varied between 7.7 - 8.9 and 8.0 - 9.0 during the Pre-monsoon and Post-monsoon seasons of the year 2011. Some of the Pre and Post-monsoon samples exceeded the World Health Organization [23] standards.

EC varied to 400 - 3190 and 340 - 2920 $\mu\text{S}/\text{cm}$, and TDS changed to 221 - 1829 mg/l and 209 - 2527 mg/l during Pre-monsoon and Post-monsoon seasons, respectively in the year 2009. EC and TDS exceeded WHO Standard, 1993 permissible limits in the years 2009 and 2011. Calcium and Magnesium ranged between 10-168 mg/l and 12.1 - 150.7 mg/l, respectively during the Pre-monsoon, and 8 - 152 mg/l and 8.5-189.5 mg/l, respectively during the Post-monsoon season in the year 2011. Most of the groundwater samples exceed [24] [25] permissible limits in Mg. The increased concentration in Mg in groundwater samples is due to soil characteristics in the District of Erode. The ionic concentration ranged between 2 - 391 ppm and 0 - 460 ppm for Sodium in the year 2009, and 21 - 978 mg/l and between 14 - 1106 mg/l for Chloride in the year 2011, respectively. Ionic concentrations were 5 - 219 ppm for Potassium in the 2009 Pre-monsoon season, and 73.14 - 793 mg/l for Bicarbonate in the 2011 Post-monsoon season, and 11 - 331 mg/l for Sulphate during the Pre-monsoon season in the Year 2011; and 2 - 145 ppm for Potassium, 94.3 - 732 mg/l for Bicarbonate, and 11 - 219 mg/l for Sulphate during the Pre-monsoon season in the Year 2011 respectively.

In this study, it was observed that pH, EC, Mg, HCO_3 , SO_4 and Cl were higher in pre-monsoon than in Post-monsoon season in the year 2009. TDS and Na were lower, but K and Ca were higher during Pre-monsoon. It indicates that the seasonal variation of cation concentrations was very conspicuous in this environment. The relative abundance of Sodium and Magnesium during the Pre-monsoon revealed the influence of rainwater on the groundwater. Chloride and Sulphate were dominated during Pre-monsoon season, but chloride showed more variations than Sulphate during. It is resulted from agricultural activity in the shallow aquifer system. Average Value of the pH for all the four seasons is falling just below the maximum value set by the WHO Standard, 1993. Average Value of the TDS for all the four seasons is falling just above the minimum value set by the WHO. Average Value of Ca, Mg, Na, SO_4 and Cl for all the four seasons are falling within the permissible value set by the WHO. pH, Electrical Conductivity, Total Dissolved Solids and Total hardness for 10 samples and its variations trends are given in Figures 2-5. The Minimum and Maximum Parameter levels for all the Physico-chemical parameters of groundwater in the study area are given in Figure 6.

3.1. Correlation between Variables

Correlation analysis is a bivariate method applied to describe the degree of relation between hydrochemical parameters. Correlation Coefficient is a measure of linear association between two variables. The correlation coefficient ranges from -1 to 1 where -1 describes a relationship where an increase in one variable is accompanied by a predictable and consistent decrease in the other, 0 describes a random or non-existent relationship and +1

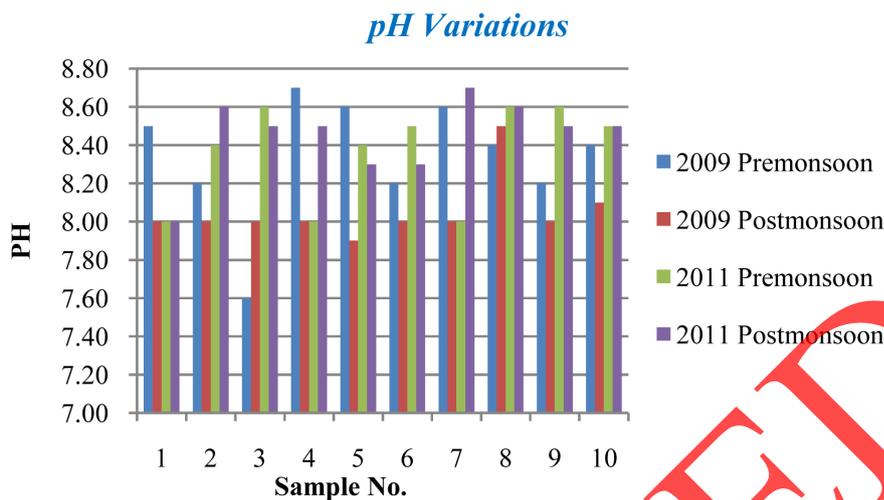


Figure 2. pH for 10 samples.

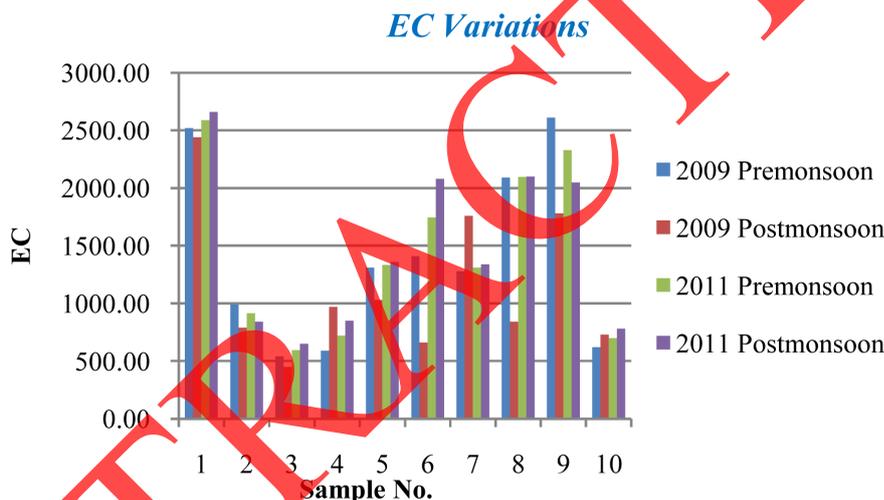


Figure 3. Electrical conductivity for 10 samples.

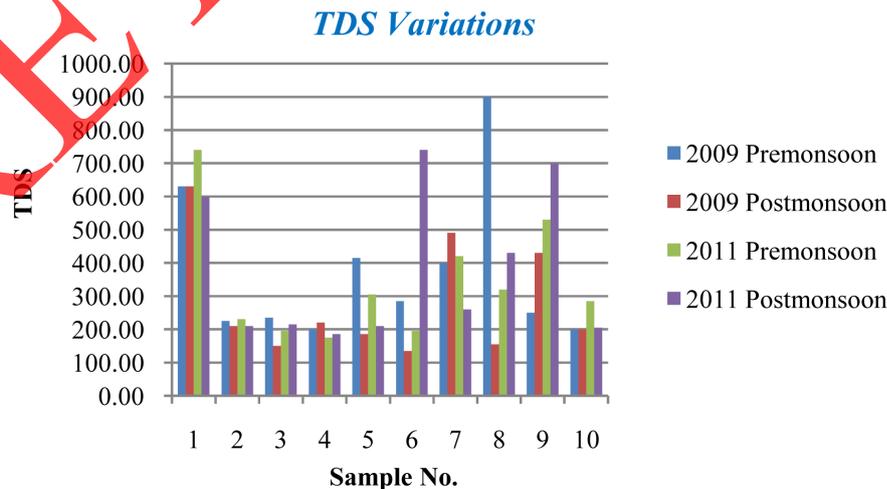


Figure 4. Total dissolved solids for 10 samples.

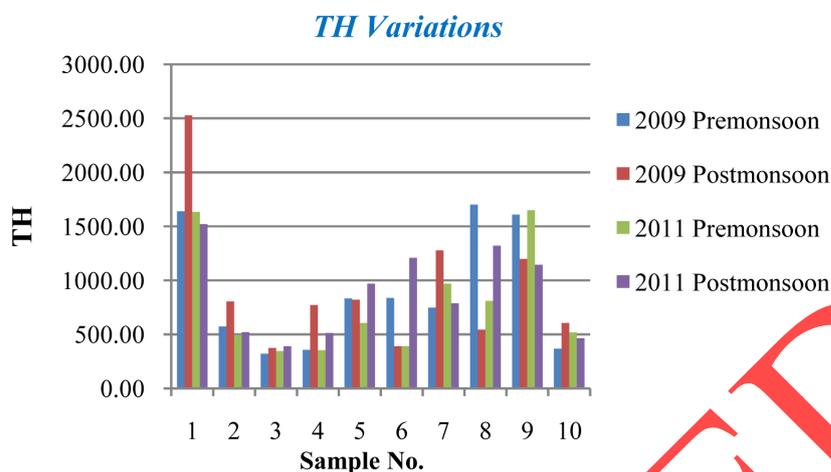


Figure 5. Total hardness for 10 samples.

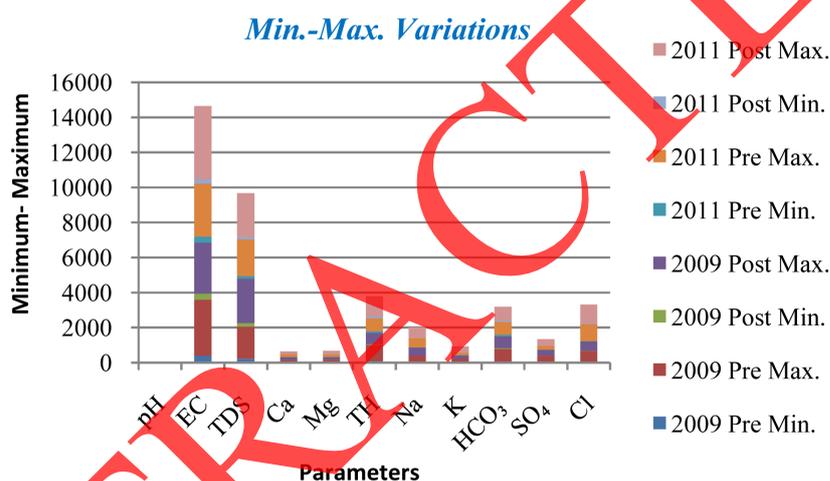


Figure 6. Minimum-maximum of parameters.

describes a relationship where an increase in one variable is accompanied by a predictable and consistent increase in the other. If one variable tends to increase as the other decreases, the correlation coefficient is negative. Conversely, if the two variables tend to increase together the correlation coefficient is positive. Correlation values of -1 or 1 imply an exact linear relationship, like that between a circle's radius and circumference. It is important to note that correlation does not imply causation. Also, a low correlation value does not mean that no relationship exists; merely that no linear relationship exists [26]-[28].

The correlation coefficients of the studied monsoon parameters are shown in **Table 2**. In the present study, Na, Cl and TDS are strongly correlated with EC. TDS is strongly correlated with Na and Cl. TH is strongly correlated with Na. Mg is strongly correlated with TH. Cl is strongly associated with TH and Na is strongly associated with Cl. pH is negatively correlated with TDS, Ca and TH. Also K is negatively correlated with Ca, Mg and TH.

3.2. Principal Components Analysis

Principal Components Analysis is used to form a smaller number of uncorrelated variables from a large set of data. The objective of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components. It is commonly used as one step in a series of analyses. Principal Components Analysis is used to reduce the number of variables and avoid multi-collinearity, as too many predictors are existing relative to the number of observations. An overview of principal component analysis can be found

Table 2. Correlation coefficients of physico-chemical parameters monsoon data 2009-11.

2009 & 2011	pH	EC	TDS	Ca	Mg	TH	Na	K	HCO ₃	SO ₄	Cl
pH	1.00										
EC	0.13	1.00									
TDS	-0.08	0.86	1.00								
Ca	-0.10	0.34	0.38	1.00							
Mg	0.01	0.64	0.64	0.38	1.00						
TH	-0.04	0.62	0.64	0.73	0.91	1.00					
Na	0.11	0.81	0.80	0.13	0.45	0.39	1.00				
K	0.08	0.36	0.41	-0.03	-0.02	-0.03	0.26	1.00			
HCO ₃	0.03	0.53	0.47	0.04	0.25	0.20	0.55	0.33	1.00		
SO ₄	0.00	0.61	0.68	0.33	0.52	0.53	0.55	0.24	0.13	1.00	
Cl	0.00	0.84	0.86	0.44	0.69	0.70	0.79	0.28	0.28	0.57	1.00

in multivariate analysis, such as [29]. 11 attributes are evaluated using PCA Method and 7 components constitute total variance of 98.3%. The first principal component has variance (Eigenvalue) 4.725 and accounts for 47.25% of the total variance. The coefficients listed under shows the principal component scores for the first four Principal Components PC1, PC2, PC3 and PC4:

$$PC1 = 0.02pH + 0.42EC + 0.24Ca + 0.37Mg + 0.38TH + 0.36Na + 0.13K + 0.21HCO_3 + 0.32SO_4 + 0.41Cl \quad (1)$$

$$PC2 = 0.24pH + 0.19EC - 0.44Ca - 0.25Mg - 0.38TH + 0.31Na + 0.49K + 0.41HCO_3 - 0.02SO_4 + 0.01Cl \quad (2)$$

$$PC3 = 0.93pH + 0.03EC - 0.07Ca + 0.12Mg + 0.06TH + 0.01Na - 0.24K - 0.23HCO_3 + 0.00SO_4 - 0.02Cl \quad (3)$$

$$PC4 = -0.08pH + 0.03EC - 0.22Ca + 0.21Mg + 0.05TH + 0.19Na - 0.62K + 0.55HCO_3 - 0.42SO_4 - 0.07Cl \quad (4)$$

The first principal component is representing pH, EC, Ca, Mg, TH, Na, K, HCO₃, SO₄, Cl because the coefficients of these terms have the same sign and are not close to zero. The second principal component has variance 1.66 and accounts for 16.6% of the data variability. This component is contrasting level of pH, EC, Na, K, HCO₃, Cl value with Ca, Mg, TH and SO₄. The third principal component has variance 0.98 and accounts for 9.8% of the data variability. Third component is contrasting level of pH, EC, Mg, Na value with Ca, K, HCO₃, SO₄ and Cl to some extent. Together, the first two and the first three principal components represent 63.8% and 73.6%, respectively, of the total variability. The Fourth Principal Component has variance 0.81 and accounts for 8.1% of the total variance. The remaining 3 principal components account for a 16.5% of the variability and are probably least important. The Principal Component coefficients are shown in Table 3.

The Scree plot provides this information visually in Figure 7. It is displaying the eigenvalues associated with a component or factor in descending order versus the number of the component or factor. A factor analysis was conducted on 11 different characteristics of water quality. This Scree plot shows that 4 of those factors account for most of the variability because the line begins to straighten after factor 4. The remaining factors account for a very small proportion of the variability and are likely unimportant.

3.3. Classification

Classification is a fundamental issue in machine learning and in statistics. It is one of the elementary problems which have an objective to learn from a given set of instances with class labels, which correctly assigns a class label to a test instance. The performance of a classifier is usually measured by its classification accuracy. It is a supervised data mining technique, with the goal of accurately predicting the class label for each item in a given dataset. A classification model built to predict class labels, from the attributes of the dataset, is known as a classifier. In data mining, Bayesian classifiers are a family of probabilistic classifiers, based on applying Bayes

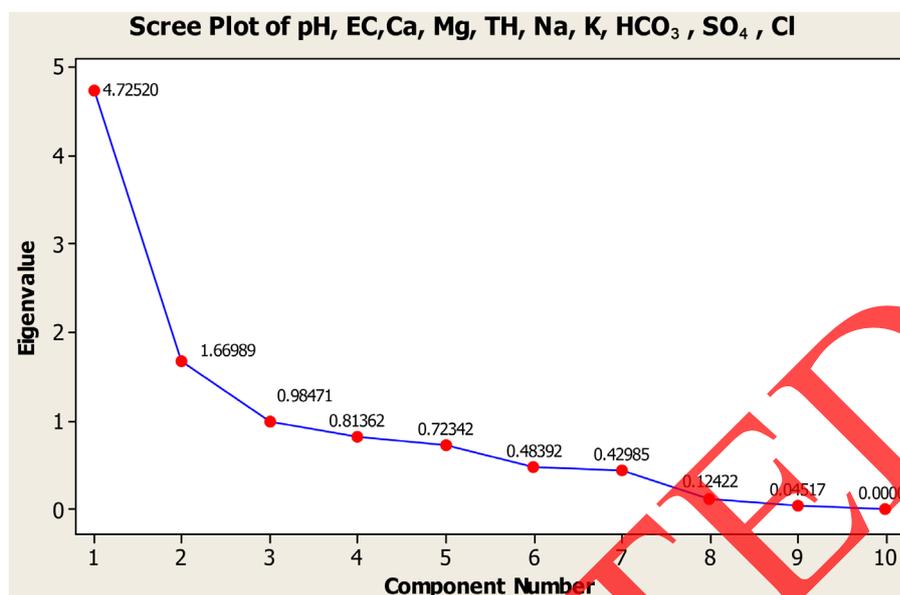


Figure 7. Minimum-maximum of parameters.

Table 3. Principal component analysis of monsoon data 2009-11.

Physicochemical Parameters	PC1	PC2	PC3	PC4	PC5	PC6	PC7
pH	0.0205	0.2437	0.9308	-0.0757	-0.2285	0.0359	-0.0668
EC	0.4211	0.1877	0.029	0.0314	0.0273	0.0693	0.1044
Ca	0.2423	-0.4394	-0.0683	-0.2171	-0.5764	0.4867	-0.0644
Mg	0.3708	-0.2515	0.1242	0.2093	0.0876	-0.6196	0.0133
TH	0.3828	-0.3817	0.0621	0.0593	-0.1915	-0.2428	-0.0186
Na	0.365	0.3069	0.0094	0.1927	0.2902	0.4067	0.1399
K	0.1312	0.4871	-0.2364	-0.6205	-0.3347	-0.3567	0.135
HCO ₃	0.2159	0.4138	-0.2293	0.5471	-0.4135	-0.0171	-0.4023
SO ₄	0.3287	-0.0239	-0.002	-0.4151	0.4269	0.0749	-0.7019
Cl	0.4176	0.0105	-0.0164	-0.0717	0.1452	0.1252	0.5364
Eigenvalue	4.7252	1.66989	0.98471	0.81362	0.72342	0.48392	0.42985
Proportion %	47.252	16.699	9.847	8.136	7.234	4.839	4.298
Cumulative %	47.252	63.951	73.798	81.934	89.168	94.007	98.306

theorem. It is a mathematical formula used for calculating conditional probabilities. It relates current probability to prior probability [30]-[32].

The most common neural network model is the multilayer perceptron (MLP). This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown. Different types of neural networks are found in literature [33] [34]. Multi-Layer Perceptron Classifier In a classification process, the outcome of MLP classifier is class membership for the given input reduct set. The advantage of a neural network algorithm is it adjusts themselves to the application by means of the training or learning process. MLP network-based classifiers have shown good results in application [35]. Multi-Layer Perceptron is feed forward neural network

which is widely using in classification of data.

WEKA Trains a multilayer perceptron with one hidden layer using WEKA’s Optimization class by minimizing the squared error plus a quadratic penalty with the BFGS method. All the attributes are standardized. There are 11 parameters. The ridge parameter is used to determine the penalty on the size of the weights. The number of hidden units can also be specified. The large numbers produce long training times. Finally, it is possible to use conjugate gradient descent rather than BFGS updates, which may be faster for cases with many parameters. To improve speed, an approximate version of the logistic function is used as the activation function. Also, if delta values in the back propagation step are within the user-specified tolerance, the gradient is not updated for that particular instance, which saves some additional time. Data is split into batches and processed in separate threads in this case.

WEKA (Waikato Environment for Knowledge Analysis) software is used for classification analysis. It is a widely adopted suite of machine learning written in Java, developed at the University of Waikato for classifier performance assessment. WEKA is free software available under the GNU General Public License. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. It is used for analyzing the selected classifiers and evaluation of module II were based on the testing set. A simple 10 fold cross validation was performed, as described in [36]. So we have used 65% for training and the remaining 35% tuples for testing. The process was repeated ten times and the accuracy for true, false, and total accuracy estimated based on confusion matrix. The final accuracy is the average of the accuracy in all tests [37]. MLP with and without PCA methods are analysed using WEKA tool and the results are listed in Table 4.

3.4. Cluster Analysis

Goal is to reduce the number of variables by combining variables with similar characteristics. Clustering of variables was done with the correlation distance measure, single linkage and dendrogram. Single Linkage (Nearest Neighbor) Method is best suited for observations or variables that are clearly separated. Dendrogram is suggesting variables which can be combined, possibly by averaging or totaling and the information is demonstrated in the form of a tree diagram.

Cluster analysis is used for grouping the cases based on the similarity of the responses to several variables. On the basis of the connecting distances between parameters and their positions on the dendrogram, distinctive clusters of the variables were defined. Cluster analysis is a useful method for combining groundwater wells into homogenous groups according to their water quality [38]. In this study, three types of clustering nature were observed during the year 2009 and 2011 before and after the monsoon seasons.

These variables would be found in natural mineralization and anthropogenic processes. Analyzing large quantity of data can be easily performed using cluster analysis. Cluster analysis deals with dividing the data into groups based on the information found on the data. In the year 1967, Johnson [39] introduced the method of Hierarchical clustering which helps in identifying “groups” or “clusters” in large group of data. In this method, the clusters are formed using different “Parameters” based on the “similarities” found on the data. The “similarity” is determined by the “Euclidean distance” between the “Parameters”. The method of clustering first assumes that all the “Parameters” considered are individual clusters by themselves. Then by determining, the “si-

Table 4. Comparison of MLP with and without PCA method.

Summary	MLP	PCA MLP
Correlation coefficient	0.9074	0.9089
Mean absolute error	122.7432	127.0982
Root mean squared error	203.0173	200.1032
Relative absolute error	32.67%	33.83%
Root relative squared error	43.31%	42.69%
Total number of Instances	224	224
Time taken to build model	0.59 Seconds	0.41 Seconds

milarities” between the “Parameters”, numbers of clusters with different “Parameters” are formed. This process of clustering different “Parameters” continues until all the “Parameters” considered together form one “strong cluster”. Since our aim is to determine the “order of influence” of the variables which affect the sustainable water quality. The hierarchical cluster analysis used to determine the hierarchical order of influence of variables on water quality. The results of the cluster analysis are shown by “dendrogram”, a cluster-tree in **Figure 8**.

Initially the cluster is formed between Mg and TH at a distance of 0.09 with a similarity level of 95.58. EC, TDS are formed as a cluster at distance of 0.13 and is again formed a cluster with Cl and Na at a distance of 0.14 and 0.19 respectively with a similarity of 93.04, 92.92 and 90.69. Ca is formed a cluster with Mg and TH at a distance of 0.27 and is again formed a cluster with Na cluster group at a distance of 0.29 with a similarity of 85.15. At a distance of 0.32 and 0.45 subsequent clusters are formed with SO₄ and HCO₃. The entire group is Cluster II. Cluster I and Cluster III is formed at a distance of 0.87 and 0.58 with a similarity of 56.39 and 70.65.

The pH quantity that forms a major part together with EC, TDS, Ca, Mg, TH, Na, HCO₃, SO₄ and Cl forms the next cluster in affecting the Water Quality in Pre-monsoon season in the year 2009 & 2011 and Post-monsoon seasons in the year 2009 & 2011. It is interesting to note that the presence of K, by linking to the HCO₃ together, means that both these have equal influence in dictating the water quality in the monsoon seasons whatever may be their quantity. The results of the Clustering analysis are listed in the **Table 5**.

4. Conclusions

This research demonstrated the hydrogeochemical evolution of shallow groundwater via multiple approaches. The groundwater quality with respect to agricultural and drinking purposes was assessed. The conclusions are summarized as follows.

Groundwater quality of an area must be studied to understand its suitability for domestic and irrigation purposes. The groundwater in this area was seen to be good and suitable for drinking and domestic purposes. Overall, groundwater in the District of Erode area remains usable. The problem of groundwater in the District of Erode is rather complicated, and is derived from excessive exploitation of the aquifer systems. The study area depends totally on the water derived from the aquifer for its irrigation use and industrial activities. The over-pumping of fresh water caused the deterioration of the water quality by the upwelling of deep saline water. Fifty-six groundwater samples were collected from the Erode district during the monsoon seasons in the years

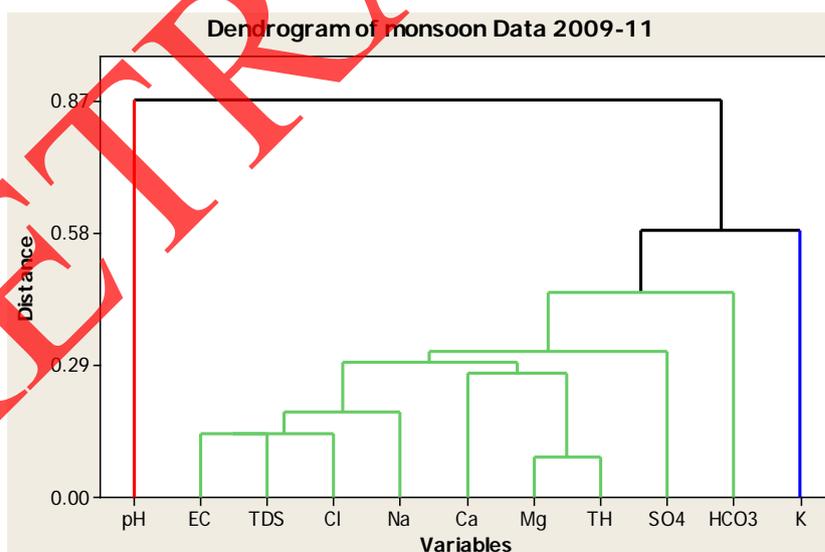


Figure 8. Dendrogram of monsoon data 2009 & 2011.

Table 5. Cluster analysis of physico-chemical parameters.

Year	Cluster I	Cluster II	Cluster III
2009 & 2011	pH	EC TDS Ca Mg T Na HCO ₃ SO ₄ Cl	K

2009 and 2011, and physico-chemical characteristics of the samples were studied to determine the quality and the suitability of the groundwater for various purposes as well as the characteristics of groundwater quality. All the groundwater samples collected from the Erode District, Tamil Nadu, India, showed that the major ions were fallen within the permissible range. Based on TDS, 69% of the groundwater was fresh and permissible for drinking. The groundwater varied from soft to very hard on the basis of TH. 27% of the groundwater is not suitable for drinking based on pH. 90% of the groundwater is suitable for drinking based on EC. Calcium level is 100% permissible in the District of Erode. Calcium level is within the range of permissible level due to regular variations in the seasonal monsoons. Total hardness is 41.5% significant in rejecting the water quality. The concentration of sodium and potassium is 81% and 83% respectively in accepting the water quality. Cl and SO₄ are 75.8% and 92.8% in accepting the water quality in the District of Erode.

The values of correlation coefficients and their significance levels contributed to the selection of proper treatments to minimize the contamination of the groundwater in the study area. The principal component analysis reveals that the groundwater from the District of Erode area has been greatly influenced by environmental events in the area. Classification result of the MLP with PCA is giving better conclusion about the Water Quality Parameters than the MLP method alone in terms of correlation coefficient and time taken to build a model. A dendrogram of 224 cases and 11 variables of monsoon seasons in the years 2009 and 2011 is plotted and grouped into three main clusters. Most water samples are acceptable for agricultural use. The suitability of groundwater with regard to agricultural use is controlled by salinity. The present groundwater quality status must be maintained by taking precautionary measures such as rainwater harvesting, less use of chemical fertilisers and the ongoing monitoring of groundwater quality in this region. A continuous monitoring program of the water quality will help to avoid further deterioration of the groundwater quality in this region. Necessary management practices are required when doubtful-quality water is used for irrigation.

References

- [1] Davis, S.N. and Wiest, D. (1996) *Hydrogeology*. John Wiley & Sons, New York.
- [2] Brydon, D.A. and Frodsham, D.A. (2001) A Model-Based Approach to Predicting BOD5 in Settled Sewage. *Water Science and Technology*, **44**, 9-15.
- [3] Chandramouli, V., Brion, G., Neelakantan, T.R. and Lingireddy, S. (2007) Backfilling Missing Microbial Concentrations in a Riverine Database Using Artificial Neural Networks. *Water Research*, **41**, 217-227. <http://dx.doi.org/10.1016/j.watres.2006.08.022>
- [4] Cho, J.H., Sung, K.S. and Ha, S.R. (2004) A River Water Quality Management Model for Optimising Regional Wastewater Treatment Using a Genetic Algorithm. *Journal of Environmental Management*, **73**, 229-242. <http://dx.doi.org/10.1016/j.jenvman.2004.07.004>
- [5] Han, J., Kamber, M. and Pei, J. (2011) *Data Mining: Concepts and Techniques*. 3rd Edition, Morgan Kaufmann Publishers, Burlington.
- [6] Içaga, Y. (2005) Genetic Algorithm Usage in Water Quality Monitoring Networks Optimization in Gediz (Turkey) River Basin. *Environmental Monitoring and Assessment*, **108**, 261-277. <http://dx.doi.org/10.1007/s10661-005-4328-z>
- [7] Ip, W.C., Hu, B.Q., Wong, H. and Xia, J. (2009) Applications of Grey Relational Method to River Environment Quality Evaluation in China. *Journal of Hydrology*, **379**, 284-290. <http://dx.doi.org/10.1016/j.jhydrol.2009.10.013>
- [8] O'Connor, E., Smeaton, A.F., O'Connor, N.E. and Regan, F. (2012) A neural Network Approach to Smarter Sensor Networks for Water Quality Monitoring. *Sensors*, **12**, 4605-4632. <http://dx.doi.org/10.3390/s120404605>
- [9] Şenkal, O., Yıldız, B.Y., Şahin, M. and Pestemalci, V. (2012) Precipitable Water Modelling Using Artificial Neural Network in Cukurova Region. *Environmental Monitoring and Assessment*, **184**, 141-147. <http://dx.doi.org/10.1007/s10661-011-1953-6>
- [10] Smeti, E.M., Thanasoulas, N.C., Lytras, E.S., Tzoumerkas, P.C. and Goulinopoulos, S.K. (2009) Treated Water Quality Assurance and Description of Distribution Networks by Multivariate Chemometrics. *Water Research*, **43**, 4676-4684. <http://dx.doi.org/10.1016/j.watres.2009.07.023>
- [11] Zou, R. and Lung, W.S. (2004) Robust Water Quality Model Calibration Using an Alternating Fitness Genetic Algorithm. *Journal of Water Resources Planning and Management*, **130**, 471-479. [http://dx.doi.org/10.1061/\(ASCE\)0733-9496\(2004\)130:6\(471\)](http://dx.doi.org/10.1061/(ASCE)0733-9496(2004)130:6(471))
- [12] Chen, T. and Chen, H. (1995) Approximation Capability to Functions of Several Variables, Nonlinear Functions and Operator by Radial Basis Function Neural Network. *IEEE Transactions on Neural Networks*, **6**, 904-910. <http://dx.doi.org/10.1109/72.392252>

- [13] Hagan, M.T., Demuth, H.B. and Beale, M.H. (2004) Neural Network Design. China Machine Press, Beijing.
- [14] Zhao, L.M., Hu, H.Y., Wei, D.H. and Wang, S.Q. (1999) Multilayer forward Artificial Neural Network. Yellow River Conservancy Press, Zhengzhou.
- [15] Luo, D.G., Guo, Q. and Wang, X.J. (2003) Neural Network Model Design of Surface Water Environmental Quality Assessment. *Geography and Geo-Information Science*, **19**, 77-81 (in Chinese).
- [16] Fu, Y.F., Zhang, J. and Luo, G.M. (2004) Application of BP Network to Groundwater Quality Evaluation. *Journal of Northwest Sci-Tech University of Agriculture and Forestry*, **32**, 129-132.
- [17] Cheng, C.T., Chau, K.W., Sun, Y.G. and Lin, J.Y. (2005) Long-Term Prediction of Discharges in Manwan Reservoir Using Artificial Neural Network Models. In: Wang, J., Liao, X.-F. and Yi, Z., Eds., *Advances in Neural Networks—ISNN 2005*, Springer, Berlin, Heidelberg, 1040-1045. http://dx.doi.org/10.1007/11427469_165
- [18] Liu, W.C. and Chung, C.E. (2014) Enhancing the Predicting Accuracy of the Water Stage Using a Physical-Based Model and an Artificial Neural Network-Genetic Algorithm in a River System. *Water*, **6**, 1642-1661. <http://dx.doi.org/10.3390/w6061642>
- [19] Chen, W.B. and Liu, W.C. (2015) Water Quality Modeling in Reservoirs Using Multivariate Linear Regression and Two Neural Network Models. *Advances in Artificial Neural Systems*, **2015**, Article ID: 521721. <http://dx.doi.org/10.1155/2015/521721>
- [20] Dalal, S.G., Shirodkar, P.V., Jagtap, T.G., Naik, B.G. and Rao, G.S. (2010) Evaluation of Significant Sources Influencing the Variation of Water Quality of Kandla Creek, Gulf of Katchchh, Using PCA. *Environmental Monitoring and Assessment*, **163**, 49-56. <http://dx.doi.org/10.1007/s10661-009-0815-4>
- [21] Olsen, R.L., Chappell, R.W. and Loftis, J.C. (2012) Water Quality Sample Collection, Data Treatment and Results Presentation for Principal Components Analysis—Literature Review and Illinois River Watershed Case Study. *Water Research*, **46**, 3110-3122. <http://dx.doi.org/10.1016/j.watres.2012.03.028>
- [22] CGWB (2008) Report Groundwater Information of Erode District. Tamil Nadu Central Ground Water Board, Chennai.
- [23] WHO (1993) Guideline of Drinking Quality. World Health Organization, Geneva.
- [24] APHA (1998) Standard Methods for the Examination of Water and Wastewater. American Public Health Association, Washington DC.
- [25] BIS (1998) Standards IS: 10500. Bureau of Indian Standards. Manak Bhavan, New Delhi.
- [26] Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez J.M. and Fernandez, L. (2000) Temporal Evolution of Groundwater Composition in Alluvial Aquifer (Pisuerga River, Spain) by Principal Component Analysis. *Water Research*, **49**, 359-372. [http://dx.doi.org/10.1016/S0043-1354\(99\)00225-0](http://dx.doi.org/10.1016/S0043-1354(99)00225-0)
- [27] Kim, R.H., Yum, B.W. and Chang, H.W. (2002) Hydrogeochemical and Isotopic Characteristics for Salinization of a Shallow Groundwater in Coastal Area, Youngkwang, Korea. *17th Salt Water Intrusion Meeting*, Delft, 6-10 May 2002, 227-237.
- [28] Holland, S.M. (2006) Cluster Analysis. Department of Geology, University of Georgia, Athens.
- [29] Harmon, H. (1976) Modern Factor Analysis. 3rd Edition, University of Chicago Press, Chicago.
- [30] Han, H.G., Chen, Q. and Qiao, J.F. (2011) An Efficient Self-Organizing RBF Neural Network for Water Quality Prediction. *Neural Networks*, **24**, 717-725. <http://dx.doi.org/10.1016/j.neunet.2011.04.006>
- [31] Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Burlington.
- [32] Petre, R. (2015) Enhancing Forecasting Performance of Naive-Bayes Classifiers with Discretization Techniques. *Database Systems Journal*, **6**, 24-30.
- [33] Benne, R.G., Dhandra, B.V. and Hangarge, M. (2009) Tri-Scripts Handwritten Numeral Recognition: A Novel Approach. *Advances in Computational Research*, **1**, 47-51.
- [34] Dhandra, B.V., Benne, R.G. and Hangarge, M. (2011) Kannada, Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network: A Script Independent Approach. *International Journal of Computer Applications*, **26**, 11-16. <http://dx.doi.org/10.5120/3134-4319>
- [35] Zhang, G.P. (2000) Neural Networks for Classification: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **30**, 451-462. <http://dx.doi.org/10.1109/5326.897072>
- [36] Cao, L., Liu, X., Wang, Z.P. and Zhang, Z. (2013) The Spatial Outlier Mining Algorithm Based on the KNN Graph. *Journal of Software*, **8**, 3158-3165. <http://dx.doi.org/10.4304/jsw.8.12.3158-3165>
- [37] Sudha, M. and Valarmathi, B. (2014) Rainfall Forecast Analysis Using Rough Set Attribute Reduction and Data Mining Methods. *Agris On-Line Papers in Economics and Informatics*, **6**, 145-154.

- [38] Venkatramanan, S., Chung, S.Y., Ramkumar, T., Gnanachandrasamy, G. and Vasudevan, S. (2013) A Multivariate Statistical Approaches on Physicochemical Characteristics of Ground Water in and around Nagapattinam District, Cauvery Deltaic Region of Tamil Nadu, India. *Earth Sciences Research Journal*, **17**, 97-103.
- [39] Johnson, S.C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, **32**, 241-254.
<http://dx.doi.org/10.1007/BF02289588>

RETRACTED