Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110519

Application of multivariate statistical techniques for sub soil water quality analysis

Ganesan Shanthi[1] *and D. R. Subasree*[2,*]

[1] Department of Civil Engineering,Shivani Engineering College,Trichirappalli, India
[2] Department of Civil Engineering, St. Joseph's College of Engineering & Technology, Thanjavur, India

Received: 22 Jul. 2017, Revised: 8 Aug. 2017, Accepted: 9 Aug. 2017
Published online: 1 Sep. 2017

Abstract: Multivariate Statistical method was used for analysis of water quality samples. Ground water quality samples were collected from Pudukkottai District, Tamil Nadu, and South India. The following parameters such as EC, pH, Ca, Mg, Na HCO3, SO4, NO3, KCl, TDS were analyzed and compared with standards. Principal Component Analysis combined with Piper tri-linear diagram, were used for assessment of data. Four principal components were derived, with cumulative variance of 74.3975%. The underlying factor which is responsible for variation in the water quality has been obtained. Piper trilinear diagram shows the graphical representation of water chemistry in the study area. It shows the most of the samples falls under Ca^{2+}-Mg^{2+}-HCO_3^- water type. All the parameters analyzed are under the desirable limit given by BIS standards hence it is suitable for domestic and irrigation purposes.

Keywords: Key Words: Ground Water Quality Analysis, Multivariate Statistical Techniques, Principal Component Analysis, Piper Trilinear Diagram

1 Introduction

Water stored beneath the ground which percolates through rocks and soil is called ground water. When the ground water is stored in rocks then it is called as aquifers, Water percolates freely through these rocks (gravel, sandstone, sand or lime stone). Movement of water between the rocks is free without restrictions, hence there is large amount of connected spaces which makes them permeable. Ground Water plays an important role for the survival of human being; it is a major source for domestic purposes. One third of world's drinking water is contributed by ground water. Quality of ground water is more significant for domestic, agricultural and industrial uses. Contamination of ground water due to infiltration of leechate from landfill sites, usage of fertilizers and pesticides, intrusion of salt water are some of the sources for pollution of ground water [1]. For sustainable domestic, industrial and agriculture usages the quality of water is an important factor. Most of the researchers in India identified the impacts of discharge of toxic substances and Effluent discharge from industries into the river [2]. Water is considered as the precious and fundamental source for life. As the surface water is highly polluted, which does not meet the standards for drinking, and also availability of surface water is deteriorating nowadays, groundwater is the dependable resource. Ground water quality is a very important problem throughout the globe. There are many factors, involved which cause changes in groundwater quality [3].

For efficient ground water quality management, assessment of water quality parameters and comparing the same with standards is not sufficient for arriving some useful conclusion. Since the Water quality data consists of more variables; some advanced approaches are needed for getting useful information. Statistical analysis of water quality data found as a useful approach for getting the hidden information from the data and without changing the original facts [4]. Data are wealth, A great quantity of data are available, but they are useful only when they are used in systematic way. Such a systematic review and interpretation of the raw data gives huge amount of information. Data set may be complex due to various causes. It may contain many observations and finalizing the variable, which causes the issue, is tough. In many cases data are measured in number of units and on every unit more number of variables are measured. To be aware of a large data set on various dependent variables, the data

* Corresponding author e-mail: subasreecivil@yahoo.com

should be summarized. If there is only single variable for under concern, they are characterized by mean, skewness, kurtosis etc. In multivariate data the calculations is simplified by matrix notations [5].

Principal component which is the prime factor for PCA is calculated as follows.
The first principal component Y_1 is calculated using the formula

$$Y_1 = wa_1x_1 + wa_2x_2 + \cdots + wa_nx_n \tag{1}$$

(or)

$$Y_1 = w_1^T X \tag{2}$$

Variance of Y_1 can be maximized, by using maximum values of weights. But it can be controlled by having the sum of the squares of weights as equal to 1.

$$w_1^2 + w_2^2 + \cdots + w_n^2 = 1 \tag{3}$$

Second Principal component (Y_2) is calculated ensuring it is uncorrelated with the previous principal component.

$$Y_2 = wb_1x_1 + wb_2x_2 + \cdots + wb_nx_n. \tag{4}$$

Further calculation continues until n number of principal components arrives and it should be equal to the original number of variables.

Conversion of principal component from original values is equal to

$$Y = PX \tag{5}$$

where P is the row of matrix and it is called as Eigen vectors. Weights forms the elements of Eigen vectors and it also called as loadings of the Eigen vector [6].

Let P is a matrix which transforms X into Y.
X is the original data taken, and Y is the re-demonstration of the data set Eq. (5) gives the transformation of the original values, where P gives rotation and a linear stretch which then transforms X into Y. The rows of P, $\{p_1, \ldots, p_n\}$, which is a set of new vectors and expresses the columns of X. Next step is the change of basis.

$$PX = \begin{bmatrix} P_1 \\ \vdots \\ P_m \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \tag{6}$$

$$Y = \begin{bmatrix} P_1 \cdot X_1 & \cdots & P_1 \cdot X_n \\ \vdots & \ddots & \vdots \\ P_m \cdot X_1 & \cdots & P_m \cdot X_n \end{bmatrix} \tag{7}$$

Let the form of column of y as

$$y_i = \begin{bmatrix} p_1 \cdot x_1 \\ \vdots \\ p_m \cdot X_i \end{bmatrix} \tag{8}$$

Multivariate statistical analysis, such as Factor analysis (FA)/Principal component analysis (PCA), Cluster analysis (CA) Canonical correspondence analysis (CCA),

Fig. 1: Pudukkottai district map.

Discriminate analysis (DA), Redundancy analysis (RDA), combined with graphical representation like Piper tri-linear diagram, Stiff diagram, Maucha diagram act as an important tool in environmental science. Factor Analysis gives the hidden information about the variables. Unclassified data is used in cluster analysis to make observational groups. PCA extracts linear relationships from variable set, it explains the variations in a large data set of correlated variables from smaller set variables [7–11].

In India ground water is a major source for both rural and urban areas. Groundwater occupies a dominant role in agriculture practices in India. Usage of ground water for agriculture has been accounted for last three decades and calculated 70% of India's food production. Ground water proves to be better quality when compared to surface water, and also gives good yield in irrigation. It provides consistent supply of water, and it is a significant one for irrigation. In the country the ground water, which occurs in shallow aquifers which is appropriate for various purposes is mainly of mixed type, calcium bicarbonate type, and other types dominates are sodium chloride type. In this study ground water quality assessment has been carried out in 21 villages of Gandarvakkottai taluk, Pudukkottai district, Tamil Nadu, south India, and Multivariate statistical Analysis combined with Piper trilinear diagram has been used for water quality.

2 Study area

Current study was carried out in Pudukkottai district, where the ground water is a major resource for drinking and agriculture. The district lies between 78° 25′ and 79° 15′ east longitude and between 9° 50′ and 10° 40′ of the north latitude. It has an area of 4,663 km^2 with coastal line of 42 kms. Terrain of Pudukkottai District is of general flatness, inter spread with small rocky hills mostly

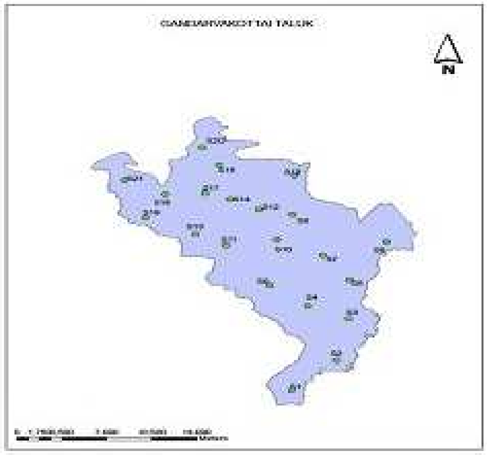

Fig. 2: Sample location from study area.

in the south western parts of the district, depressions and slopes found in the terrain are useful to construct tanks for irrigation purpose.

Average annual rainfall in the study area is 821mm, it falls under the tropical region the temperature is very high in this region, study area comes under 58 J/9, 10, 11, 14, 15, 16, 58 N/2, 3, 4 and 58 O/1& 2 of Survey of India (SOI) topographic sheets of 1:50,000 scale [12].

3 Materials and methods

Samples were collected from 21 villages in Gandarvakottai taluk of Pudukkottai district, and sampling points were shown in Fig. 2 The ground water samples are collected in 2-litre clean polyethylene bottles, before sampling bottles were cleaned well with distilled water. Standard methods were followed for analysis of water (APHA 1995). The pH was measured by digital pH meter, Total dissolved solids by gravimetric method. Digital conductivity meter was used to measure electrical conductivity (EC). Sodium was measured by flame photometer. The remaining parameters were analyzed by titrimetric methods.

4 Results and discussion

4.1 Water Quality

Statistical analysis of water quality data are shown in Table 1 All the samples taken were under permissible limit as per IS 10500–1991.

4.2 Principal Component Analysis (PCA)

PCA it is a process which gives the linear combination of the original data in order to produce the axes which is called as Principal components. Principal components are the variables which form the linear arrangement of the accessible variables. These variables are explained to report the hidden data structure, by classifying the highly correlated data in a well organized manner; It is a data reduction method by searching elemental variables which are reproduced in the observed variables. Let the given data matrix has take n variables and N samples, analysis of samples starts with centering on the means of each and every variables.

Table 1: Statistical analysis of water quality data.

Variables	min	max	Median	Mean	SD
EC	150	1500	490	516.67	340.28
pH	7.0	7.30	7	6.74	1.39
Ca	10	40	20	23.05	9.89
Mg	9	73	21	22.57	15.29
Na	2	161	39	46.19	44.014
K	1	39	1	5.95	10.47
HCO3	43	458	98	144.43	103.70
SO4	1	41	11	13.62	10.72
Cl	14	230	53	70.38	57.08
NO3	1.10	68	19	21.77	18.02
TDS	96	960	275.2	320.15	221.85

First step in PCA is finding out the correlation coefficients, it gives the strength between each variable and its relationships with one and another [13, 14]. Correlation coefficient is calculated by Karl Pearson coefficient of correlation method. It is a measure of intensity or degree of linear relationship between two variables, and it is a numerical measure of linear relationship between the two variables. The variables x and y under study are linearly related and each of the variables which is being affected by a large number of independent contributory causes of such a nature to produce a normal distribution.

$$r = \frac{N\sum xy(\sum x)(\sum y)}{\sqrt{[N\sum x^2-(\sum x)^2][N\sum y^2-(\sum y)^2]}} \tag{9}$$

where r is Karl Pearson correlation coefficient, N is the number of samples and x, y are variables.

Correlation Matrix in factor analysis gives the correlation between the ground water variables. From range 0.99 to 1.00 it represents ideal correlation, when correlation coefficient ranges from 0.80 to 0.98 it represents the powerful correlation.

When it ranges from 0.5 to 0.8 it is considered as fair correlation. If the correlation coefficient value is less than 0.5 it gives the weak relationship between the parameters. when the correlation coefficient is negative it represents the negative relationship between the chemical parameters.

Correlation coefficient for each and every sample was calculated and the correlation matrix has been shown in Table 2. Strong correlation between the variables (EC and TDS), (EC and Na), (EC and Mg) was found, high correlation found between TDS and Na, Mg and HCO_3, which shows the relatively more presence of dissolved

Table 2: Correlation matrix.

	EC	pH	Ca	Mg	Na	K	HCO3	SO4	Cl	NO3	TDS
EC	1										
pH	0.090	1									
Ca	0.785	0.048	1								
Mg	0.927	-0.020	0.694	1							
Na	0.926	0.138	0.599	0.774	1						
K	0.154	0.107	0.274	-0.056	0.262	1					
HCO3	0.879	-0.024	0.738	0.834	0.804	0.171	1				
SO4	0.643	0.102	0.343	0.622	0.615	-0.003	0.311	1			
Cl	0.862	0.163	0.602	0.784	0.816	0.129	0.744	0.529	1		
NO3	0.068	0.288	-0.049	0.068	0.031	0.108	-0.183	0.473	0.255	1	
TDS	0.977	0.072	0.717	0.893	0.958	0.128	0.891	0.593	0.822	-0.026	1

Table 3: Rotated Component Matrix.

	Component			
	1	2	3	4
EC	.393	.672	.152	-.203
pH	.664	.077	.569	-.046
Ca	.300	-.095	-.079	.885
Mg	.142	-.272	-.430	-.503
Na	.278	.704	.386	-.246
K	.232	-.793	.201	-.129
HCO3	-.007	.693	-.118	.467
SO4	.872	.087	-.044	.060
Cl	.909	.049	-.011	.013
NO3	.777	-.055	.300	.228
TDS	.098	-.078	.847	-.013

Table 4: Factor Extraction using Principal component Analysis.

Component	Initial eigen values			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.460	31.456	31.456	2.869	26.079	26.079
2	2.000	18.181	49.637	2.210	20.091	46.169
3	1.521	13.823	63.460	1.586	14.419	60.588
4	1.203	10.937	74.397	1.519	13.809	74.397
5	.914	8.306	82.703			
6	.704	6.396	89.099			
7	.507	4.613	93.713			
8	.361	3.283	96.996			
9	.196	1.786	98.782			
10	.078	.706	99.488			
11	.056	.512	100.000			

salts when compared with other parameters in the samples.

Second step is Factor extraction, which is done by taking the factors with eigen value greater than 1, eigen value gives the variation of the variables which is related by that particular factor. By using orthogonal transformation it converts observation data into linearly uncorrelated values which is called as principal components. The first component indicates the most variation in the data, than the subsequent components.

In this study four factors were extracted and it is rotated by Varimax with Kaiser Normalization for achieving rotated component matrix, and it is shown in Table 3 and Table 4. Factor analysis has been carried out by SPSS 16.0 the first factor has the percentage of variance 26.079% with most positive values, and it represents the strong relationship between the other parameters, second factor has the percentage of variation of 20.091. The third factor has the percentage of variance of 14.419, The Fourth factor has the percentage of variance as 13.809 [15]. Cl, Na, TDS, Ca has the highest positive values in the I, II, III, IV components.

Hence the above said parameters are responsible for the variations in the ground water quality. The variation in

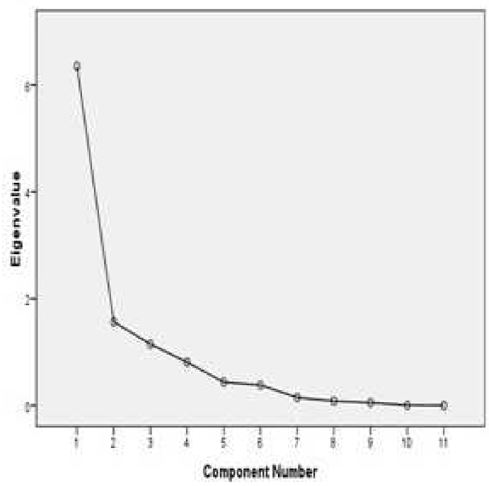

Fig. 3: Scree plot.

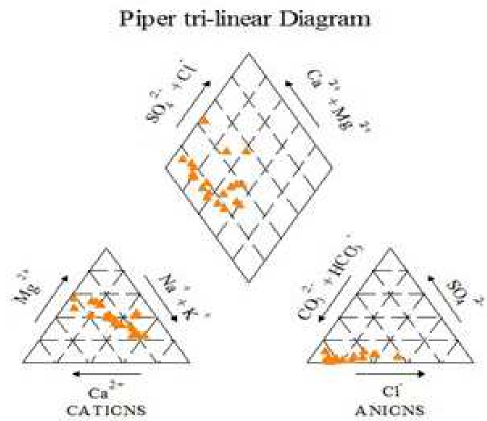

Fig. 4: Piper tri-linear representation of anions and cations.

parameters has been shown in Scree plot in Fig. 3. From the scree plot it is evident four factors are responsible for slight changes in water quality.

4.3 Piper tri-linear diagram

Piper diagram is a graphical tool which gives the chemistry of water sample. It consists of three parts in which bottom left represents the cations and bottom right represent the anion and the diamond shaped plot represent the combination of both anion and cation. Ground Water data lies at the top of the diamond convey permanent hardness. Data which lies near the left corner and shows that it is of temporary hardness. Water located at the lower corner of the diamond represents alkali carbonates. Water data arranged at right-hand side represents saline water. Mg^{2+}-HCO_3^- type hence it is considered as mixed type water.

Considering the major cations and anions in the water sample,Piper tri-linear diagram was plotted for classifying the water with respect to its anionic and cationic composition [16, 17]. Most of the samples falls in the Ca^{2+}-in their corresponding four component factors. From Piper Tri-linear diagram the water type in the study area has been categorized.

5 Conclusions

Significant information's were derived using Principal component Analysis combining with Piper Tri-linear diagram. According to (PCA), Parameters such as Cl, NaTDS, Ca are responsible for variations in the water quality. These parameter gives the highest positive loadings in their corresponding four component factors. From Piper Tri-linear diagram the water type in the study area has been categorized.

Ground water samples were analysed using standard methods (APHA). 11 parameters were evaluated from the study area and compared with the standards. Some of the samples are slightly alkaline due to the hard rock interaction in the study area. All the parameters are within the permissible limits, prescribed by the BIS Standards and it is safe for drinking and agricultural activities.

References

[1] Manoj K., Ghosh S. and Padhy P.K., Characterization and Classification of Hydrochemistry using Multivariate Graphical and Hydrostatistical Techniques Research Journal of Chemical Sciences, **3(5)**, 32–42, (2013).

[2] Sarojini Devi Boominathan, Suganthi Palanisamy, Suganthi Kanagaraj and Govindaraju Munusamy, Mapping of Ground Water Quality for Ramanathapuram Taluk of Tamil Nadu Using Geographical Information System. Cloud Publications International Journal of Advanced Remote Sensing and GIS, **4(1)**, 953–959, (2015).

[3] Mohammad N. Almasri, Mac McKee and Jagath J. Kaluarachchi Applicability of statistical learning algorithms in groundwater quality modeling Water Resources Research, **41**, (2005).

[4] Narsimha. A, Sudarshan. V, Srinivasulu. P, Venkateswarlu. P and Rajeswara Reddy. Hydrochemical concept of groundwater in and around Atmakuru area,Anantapur District, Andhra Pradesh, India, Pelagia Research Library Der Chemica Sinica, **4(1)**, 32–35, (2013).

[5] Qingchun Yang, Jianing Zhang, Yanli Wang, Yanna Fang, Jordi Delgado Martín, Multivariate Statistical Analysis of Hydrochemical Data For Shallow Ground Water Quality Factor Identification in A Coastal Aquifer Pol. J. Environ. Stud., **24(2)**, 769–776, (2015).

[6] Guey-Shin Shyu, Bai-You Cheng, Chi-Ting Chiang, Pei-Hsuan Yao and Tsun-Kuo Chang, Applying Factor Analysis Combined with Kriging and Information Entropy Theory for Mapping and Food Technology, **6(1)**, (2013).

[7] Manish Kumar, Yashbir Singh Interpretation of Water Quality Parameters for Villages of Sanganer Tehsil, by Using Multivariate Statistical Analysis J. Water Resource and Protection, **2**, 860–863, (2010).
[8] Mohammad Ali Zare Chahouki Multivariate Analysis Techniques in Environmental Science, Earth and Environmental Sciences.
[9] Rezwanul Mahmud1, Naoto Inoue1 and Ranjit Sen2 Assessment of Irrigation Water Quality By Using Principal Component Analysis In an Arsenic Affected Area of Bangladesh, J. Soil. Nature. **1(2)**, 8–17, (2007).
[10] Steven M. Holand Principal Components Analysis (PCA) Department of Geology, University of Georgia, Athens, GA 30602-2501
[11] Jon Shlensa Tutorial On Principal Component analysis Derivation, Discussion and Singular Value decomposition, 25 March 2003.
[12] Sumathi, M, Kumarasamy, K, Thiyagarajan, M and Punithavathi, A Study On The Temperature and Rainfall Conditions of Pudukkottai, District, Tamil Nadu, India International Journal of Development Research, **1(3)**, 22–26, (2011).
[13] Qingchun Yang, Jianing Zhang, Yanli Wang, Yanna Fang, Jordi Delgado Martín, Multivariate Statistical Analysis of Hydrochemical Data For Shallow Ground Water Quality Factor Identification in A Coastal Aquifer Pol. J. Environ. Stud., **24(2)**, 769–776, (2015).
[14] Narmatha. T,Jeyaseelan. A, Mohan. S.P, Ram Mohan. V, Integrating Multivariate Statistical Analysis with GIS for Groundwater in Pambar subbasin, Tamil Nadu, India International Journal of Geomatics and Geosciences, **2(2)**, (2011).
[15] Asaad M. Armanuos, Abdelazim Negm, and Oliver C. Saavedra Valeriano Groundwater Quality Investigation Using Multivariate Analysis—Case Study: Western Nile Delta Aquifer, Egypt International Journal of Environmental Science and Development, **7(1)**, (2016).
[16] P.J Sajil Kumar Interpretation of groundwater chemistry using piper and chadha's diagrams: A comparative study from Perambalur taluk, Geoscience, **54**, 12208–12211, (2013).
[17] Shimaa M. Ghoraba and A.D. Khan, Hydrochemistry and Groundwater Quality Assessment in Balochistan Province, Pakistan, IJRRAS, **17(2)**,

Ganesan Shanthi, M.E., Ph.D, Principal, Shivani Engineering College, Tiruchirappalli, received her Bachelors degree in Civil Engineering and Post Graduate degree in Water Resources Engineering and Management, from Regional Engineering College, Bharathidasan University. She has received her PhD Degree from Anna University .She has around 29 years of rich experience (17 years of Teaching experience in Engineering Colleges, 2 years of Industrial experience and 10 years, as Principal in Engineering colleges). She has served as the Member of Board studies in Engineering for Bharathidasan University, Trichy, Tamil Nadu for the period of 3 years, currently she is the member of Academic council Anna University Chennai for a period of 3years. She has published more than 15 research papers in various International, National Journals and Conference Proceedings. She has presented a paper in the 7th International Conference on Hydro Informatics, HIC 2006, at NICE, Sophia antipolis at France. She is a recognized research supervisor of Anna University. Her area of interest includes Derivation of Reservoir Operating Policies using Genetic Algorithms, Ground water Quality Analysis and Modelling, Water Shed Management, etc.

D. R. Subasree received her Bachelors degree in Civil Engineering and Post Graduate degree in Environmental Engineering from Periyar Maniammai College of Technology for women Thanjavur. She is currently doing her PhD under the Faculty of Civil Engineering in the area of Water resources from Anna University Chennai. She has around 11 years of teaching experience in various engineering colleges. Presently she is serving as an Assistant Professor and Head in the Department of Civil Engineering at St. Josephs College of Engineering and Technology, Elupatti, Thanjavur. She has published more than 8 research papers in various International Journals and Conference Proceedings. Her area of interest includes Watershed Management, Automation in Water treatment Process, Ground Water Quality Modeling, and Waste water treatment etc.