Water quality trends and predictive analysis using machine learning techniques: A python based approach

Shankar B.S[1] , Vijayalaxmi Yalavigi[2]

[1] Dept. of Civil Engineering, Cambridge Institute of Technology,
Bangalore, Karnataka, India

[2] Dept. Information Science and Engineering, Cambridge Institute of Technology,
Bangalore, Karnataka, India

*Corresponding author[1]: shanky5525@gmail.com

[1]ORCID ID: 0000-0002-8384-0771

Abstract

The perils of consuming contaminated groundwater are well documented. The present study deals with the quality assessment of the ground waters of Peenya industrial area, Bengaluru, India. The study has been carried out using machine learning techniques of python programming for 30 groundwater samples collected in and around the area, during two seasons (pre- and post-monsoon seasons for the years 2017, 2018 and 2019, and analyzed as per the protocols of American Public Health Association. The present work has been carried out in three phases. The first phase involves the coarse-grain level analysis of water quality components year wise for three years using Principal Component Analysis (PCA). The second phase involves fine-grain analysis for individual critical water quality components and the third phase involves forecasting of the water quality components based on the three- year analyses carried out in this study using time-series analysis models of machine learning. The analysis results revealed that the study area was highly polluted as seen by the non-potability of 76.67% of the samples due to the presence in excess of one or more parameters such as nitrates, hardness, total dissolved solids, calcium, magnesium, pH, fluoride, chloride, iron and chromium. The statistical findings revealed that there is an increasing tendency of the water quality parameter concentrations, though the variations are marginal. In extension with the traditional test, the statistical and predictive analysis approach has aided in revealing the parameters that influence water quality.

Keywords: Data mining techniques; groundwater; pollution; principal component analysis; water quality.

1. Introduction

1.1 Groundwater quality and monitoring

One of the most vital resources for the benefit and survival of mankind is ground- water, which along with the ice caps happens to be the largest receptor for freshwater. As per the latest abstraction, groundwater withdrawal represents almost one-fourth of the total fresh water across the globe, which aids to supply close to 50% of drinking water and close to 43% for irrigation (Gun, 2012). One of the top-most priorities to be considered is to protect our precious resource-the groundwater, not just from health threat point of view, but also for the conservation of eco system and food supplies maintenance. It is in this connection that a worldwide quality assessment of groundwater is absolutely essential because of the continued huge pressure exerted by anthropogenic influence and the drastic variations in climate. Several countries deem it that keeping the groundwater naturally clean is the best way to go, rather than any advanced treatment, which could be extremely expensive. Thus, it becomes highly imperative to understand the source of pure groundwater storage, in addition to comprehending the continuous threats to this precious resource of ours (WWQA, 2021).

It is very evident that a huge number of natural as well as anthropogenic contaminants. in addition to the bacterial contaminants enter the aquifers and this is a worldwide phenomenon. Recent evidence stresses on more attention being required to reduce contamination adjacent to well head (Lapworth et al., 2020; Ravenscroft et al., 2017) .The wide range of variations in the quality of groundwater and their behavior either isolated or in combination with other parameters/pollutants necessitate a high level of expertise. Further, in comparison to surface waters, the quality of groundwater is very hard to comprehend and subsequently mitigate the same.

All the above factors compel that a systematic groundwater quality monitoring programme is extremely crucial and needs to be aimed and designed in accordance with the monitoring objective, which relates to tracing of some specific groundwater contaminants and subsequent remediation, short and long term campaigns and monitoring programmes to get an idea about the threats of local contaminants as well as identify the overall patterns and long-term groundwater quality trends.

1.2. Machine learning in Civil Engineering

Machine learning has rapidly diversified and become invaluable, and amongst the long list of Civil engineering problems in the domains of water resources/hydrological modelling, construction engineering and management, coastal/marine engineering, geotechnical engineering, in addition to more challenging issues such as river flow forecasting, modelling compressive strength of concrete, drought forecasting, modelling evaporation, ground water level forecasting and so on (Deka, 2019). The principles of machine learning techniques are classified under supervised and unsupervised learning. Supervised techniques are used if the data have target variable to be studied. Some of the example techniques are decision tree, support vector machines, linear regression etc. These techniques are used in predicting target variable in the dataset. Unsupervised machine learning techniques are used for descriptive analytics of the data. Principal Component Analysis (PCA) is a technique of dimensionality-reduction (DR) that is mostly used to reduce a large set of variables into a smaller one that still contains much of the details in the large set (Hasan & Abdulazeez, 2021). This technique is also used to for the classification of quality of groundwater (Mahapatra et al., 2012). Q-mode principal component analysis has been applied to classify the water samples into four different categories considering parameters such as pH, DO, turbidity, TDS, hardness, calcium ion (Ca^{++}), chloride ion (Cl^{-}), BOD, iron (Fe^{++}), sulphate (SO_4). Many researchers have worked in the area of water quality assessment. Data mining techniques like Support Vector Machine (SVM), Naïve Bayes (NB), K- Nearest Neighbour (KNN) and

Classification Based on Association Rule (CBA) have been used to predict groundwater in Jordan (Aburub & Hadi, 2016). The state of an ecosystem is dependent simultaneously on many factors and parameters; these systems are multivariate in nature. Multivariate statistical methods are employed for analysis of the water quality. Factor analysis is also one of the multivariate statistical methods which is used to find the correlation between the parameters of data (Hulya & Hayal, 2017). In another study, linear regression model is used to identify the important parameters to predict the overall water quality (Hamid et al., 2016). Water quality can also be monitored using wireless sensor networks where sensors are used to measure the quality parameters (Kofi et al., 2017) .The latest work includes use of Artificial Neural Networks (ANN), a machine learning method, to monitor and predict the water quality parameters (Gasim et al.,2021).

In this study, PCA is used to visualize the change in the water quality parameters for the three years. Change in the correlation of the water quality components each year has also been studied. Fine-grain analysis has been performed on the important parameters of the water that affect the quality of the water. The parameters under the consideration in this paper are total hardness (TH), Calcium (Ca), Magnesium (Mg), Iron (Fe), Chloride (Cl), Nitrate (NO_3), Total dissolved solids (TDS), Fluoride (F), Chromium (Cr) and pH.

2. Materials and Methods

2.1. Details of Peenya Industrial Area (study area)

“Bangalore City lies between north latitude 12° 52′ 21″ to 13° 6′ 0″ and east longitude 77° 0′ 45″ to 77° 32′ 25″, approximately covering 400 km^2 of the area (Shankar & Usha, 2021). The Peenya Industrial Area is located on the 57 H/9 Toposheet, Survey of India”. “It covers about 9 km^2 lying in the heart of Bangalore City “to the northern part and comprises of almost 2000 industries, out of which, industries such as pharmaceutical, chemical, metal plating, and leather dominate”. But “people in this area have been using polluted water for washing of clothes and utensils, cleaning, as well as several other domestic chores. The author on holding discussions with the public of PIA and also with the authorities of primary health Centre received crucial information about a number of people in this area suffering from extreme skin problems such as boils, rashes, itching sensation on their hands and legs along with experiencing severe joint pain in their hips and knees after using the water” (Shankar and Balasubramanya. 2008). Figure 1 depicts the study area (PIA), showing the sampling locations.

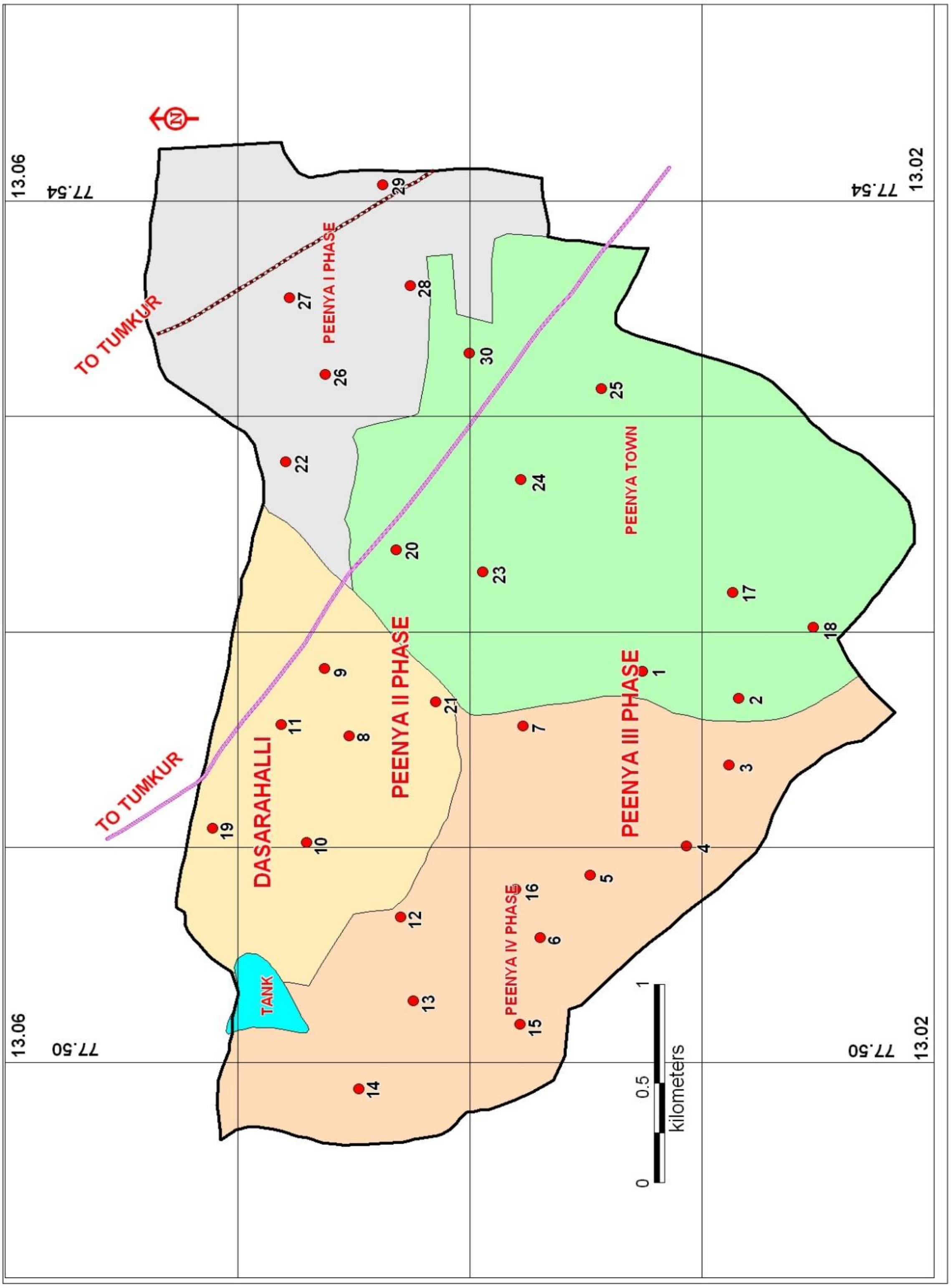


Figure. 1. Location map of Peenya Industrial Area with Sampling Stations

2.2 Sampling and analysis protocols

Thirty groundwater samples in the study area during May (pre-monsoon) and October (post-monsoon) seasons of three successive years 2017, 2018 and 2019 were collected in 2 L containers, preserved, and stored as per (APHA ,2002). The location of each station was recorded by determining the latitude and longitude of these stations using the global positioning system (GPS).

The samples were analysed for the major physicochemical parameters in the laboratory. The characteristics were determined as per the Standard methods for examination of water and wastewater (APHA 2002). The results obtained were evaluated in accordance with the standards prescribed under ‘Indian Standard Drinking Water Specification IS 10500’ of Bureau of Indian Standards, (BIS, 2012).

The analysis techniques used for the determination of water quality parameters has been presented in Table1.

Table 1. Analysis methods employed for parametric evaluation.

Water quality Parameter	Analysis methods
Chloride	Argentometric titration method
Sulphate	Spectrophotometry
Electrical Conductance	Conductivity meter
Total hardness	EDTA titrimetric analysis
Calcium	EDTA titrimetric analysis
Magnesium	Calculation though EDTA titrimetric analysis
Total alkalinity	Titration method (Neutralisation with acid)
Fluoride	UV-Visible Spectrophotometry
Nitrate	UV-Visible Spectrophotometry
Trace metals (Fe, Pb, Cu, Cr)	Atomic absorption Spectrophotometry
Sodium and potassium	Flame photometer
pH	Systronics Digital pH meter
Turbidity	Nephelo turbid meter
Acidity	Titration method (Neutralisation with base)

2.3. Methods of statistical analysis

The first phase of the work uses Principal Component Analysis to perform course- grain analysis of the water quality parameters over the years. Since the study data is multidimensional without the target variable, PCA is one of the powerful unsupervised feature extraction methods. PCA is a data analysis tool that is usually used to reduce the number of variables of a large number of interrelated variables, while retaining as much of the information (variation) as possible. PCA calculates an uncorrelated set of variables (Hamed & Reda, 2019). These components are a weighted average of the original variables given by the equation (1).

$$PCi = w_1x_1 + w_2x_2 + . . . + w_nx_n \qquad (1)$$

where w_i and x_i are the weights and original variable respectively. PCA consists of following steps i) Standardization of the variables ii) calculation of covariance matrix iii) Calculation of Eigen values and Eigen Vectors of the covariance matrix. iv) Sort the Eigen vectors from highest to lowest and select the number of components based on the largest Eigen values. v) Projection of original data to the new feature space (Salema & Hussein, 2019). The method implemented in the present work was carried out using python programming language. sklearn library provides plenty of methods for performing machine learning tasks. The data is imported in the data frame and standardized using StandardScaler method of sklearn library. The dimensions of the data are reduced using the PCA module, which considers the number of components as a parameter. The attributes of the PCA module such as explained variance_ and components are used to get the amount of variance contributed by each principal component and loadings (correlation coefficients) respectively (Buitinck et al., 2013). The scree plots help to visualize the amount of variance against the number of principal components. Factor Analysis is another statistical technique used to study the control factors of the water quality.

The second phase is fine-grain analysis, which considers the key water quality parameters for study. The water quality is more affected by TH, Ca, Mg, Fe, Cl, NO_3, TDS, F, Cr and pH. This phase studies the variations of these components each year by performing descriptive analytics. This was further aided by the first phase analysis results.

The third phase uses Time-Series analysis for forecasting the trend in the water quality parameters. The quality of the Hor Rood River was studied at Kakareza station using time series analysis in 2014. The quality of the river water was forecast using ARIMA, Auto Regressive, Integrated, Moving Average model (Tizro et al., 2014). In another study, water quality parameters were estimated using remote sensing data. Prediction of water quality was done using ARIMA (Elhag et al., 2021). Time series analysis was also used in stock price prediction. Time series analysis was combined with deep learning models to predict the stock prices (Sidra & Jaydip, 2020). In the present work, python library called Kats (Kit to analyze Time Series) was utilized for forecasting the water quality parameters. It is an open-source library developed by researchers at Facebook. It can be used to for univariate and multivariate analysis and also handle the outliers and can be used for forecasting. Kits includes many models such as Linear, Quadratic, Prophet, ARIMA etc. In the present study, Prophet model of forecasting has been employed.

3. Results & discussions

3.1 Results of physico-chemical analysis

Thirty groundwater samples in the study area were collected during May (pre- monsoon) and October (post-monsoon) seasons of three successive years 2017, 2018 and 2019 and a comprehensive analysis was carried out for 20 physico-chemical parameters for all the selected 30 groundwater stations of PIA. Since the post-monsoon analysis showed slightly higher concentrations compared to pre-monsoon, the sample results of the physicochemical groundwater analysis for post-monsoon for the year 2017 has been presented in Table 2. The analysis results revealed that the study area was highly polluted as seen by the non-potability of 76.67% of the samples due to the presence in excess of one or more parameters such as nitrates, hardness, total dissolved solids, calcium, magnesium, pH, fluoride, chloride, iron and chromium. Out of these, nitrate and total hardness accounted for 40 % and 50 % of non-potability.

A nitrate level beyond 45 mg/l is found to cause a number of health disorders, with methemoglobinemia or blue babies in infants being the most hazardous one. Apart from being carcinogenic in some cases, it was shown that women who consume groundwater with high nitrate concentrations during pregnancy are at higher risk of having a child with congenital abnormalities (Mohammadpour et al., 2022).

The maximum permissible limits for total hardness as per BIS in drinking water is 600 mg/l Groundwater with high total hardness is likely to affect human health, such as heart disease and kidney stones (Giao et al.,2022).

The main source of chromium occurs from electroplating, tanning, and leather industries. Cr concentrations greater than 0.05 mg/l can be extremely toxic and causes serious health effects (Christina et al., 2022).

Table 2. Physico-chemical analysis results for 2017 (Post-Monsoon)

Si. no	T.H	Ca	Mg	Na	K	Fe	HCO_3	CO_3	Cl	NO_3	SO_4	PO_4	TDS	EC	F	Cu	Pb	Cr	pH	Turbidity NTU
1	1094	212	124	65	1.8	1.24	302	0	342	20	628	1.0	1550	2480	2.94	0	0	0	7.72	2.1
2	633	134	65	54	2.3	1.15	206	0	238	77	302	0.5	980	1680	1.42	0	0	0	6.08	nil
3	436	90	46	92	1.4	0.00	275	0	222	35	176	0.4	800	1330	0.72	0	0	0	7.43	0.8
4	231	52	22	64	0.5	0.00	70	0	60	17	188	0.3	440	680	2.2	0	0	0	8	0.4
5	715	144	78	212	2.5	0.25	342	4	390	84	426	1.4	1510	2400	1.25	0	0	0	7.91	2.6
6	835	127	116	188	4	0.22	404	2	580	108	210	1.1	1540	2500	0.65	0	0	0	7.98	0.3
7	1260	225	155	330	3.6	0.20	532	15	700	332	203	2.1	2230	3590	0.78	0	0	0	7.94	4.1
8	756	122	101	55	5	0.20	366	0	300	101	76	0.4	940	1420	0.4	0	0	0	7.8	3.0

9	555	116	58	70	1.4	0.00	196	0	422	56	30	1.0	860	1380	0.6	0	0	0	5.1	0.1
10	476	112	42	88	3.1	0.00	198	1	264	58	122	0.5	790	1200	2	0	0	0	6.19	0.4
11	307	81	22	50	3	0.00	163	0	154	10	58	0.3	460	710	1.38	0	0	0	6.4	2.5
12	3082	512	402	100	0.4	0.00	244	0	2108	157	525	7.0	3940	6280	5.9	0	0	3	6.1	12.0
13	3110	582	366	78	4.1	0.10	370	3	1824	120	366	4.0	3530	5700	6.09	0	0	0	7.12	5.2
14	537	66	84	104	5	0.00	188	0	456	32	44	1.1	890	1420	0.63	0	0	0	7.1	2.1
15	1246	224	152	175	4	0.38	358	16	618	32	602	4.0	2010	3220	0.44	0	0	0	7.02	0.0
16	1073	218	116	145	2.3	0.40	264	0	488	18	525	2.0	1650	2640	0.39	0	0	0	7.22	0.7
17	440	118	30	80	2	0.10	275	0	260	22	112	1.1	770	1230	1.52	0	0	0	7.84	0.0
18	297	72	25	86	3	0.16	204	10	118	34	60	2.0	510	840	0.24	0	0	0	7.25	1.0
19	400	128	15	45	6	1.49	80	8	86	51	226	1.5	610	1000	1.38	0	0	0	7.7	2.0
20	459	86	54	78	4	0.05	355	0	180	25	156	3.0	760	1180	0.38	0	0	0	7.51	0.0
21	515	104	56	89	3.4	0.81	203	4	374	15	104	1.0	850	1360	1.4	0	0	0	8.22	1.4
22	648	148	60	58	1	0.00	350	4	332	19	59	0.1	860	1350	1.8	0	0	0	8	1.0
23	297	75	23	117	6	0.43	204	0	254	94	55	0.3	730	1130	0.76	0.22	0	0	6.86	1.0
24	1543	344	148	196	5	0.52	343	0	386	32	888	1.6	2170	3490	0.44	0	0	0	7.02	10.0
25	1427	275	163	282	6	0.24	316	0	876	54	842	2.0	2660	4280	1	0	0	0	7.57	0.9
26	421	106	33	107	1	0.00	394	0	180	72	86	0.0	780	1250	1.3	0	0	0	8.24	1.0
27	112	21	13	51	0.5	0.50	100	2	98	12	34	0.4	290	460	2.36	0	0	0	7.01	0.8
28	518	125	44	68	0.5	0.14	318	3	172	21	10	0.0	600	960	1.44	0	0	0	6.55	2.4
29	500	98	56	86	1.3	0.00	386	0	244	30	80	0.3	790	1220	1.41	0	0	0	6.9	0.6
30	126	35	8	62	0.6	0.08	131	0	70	12	22	1.0	280	440	1.3	0	0	0	7.1	0.0

p^H: unitless, EC in µs/cm, Turbidity in NTU, all other parameters in mg/L

3.2. PCA analysis and results

The present work was aimed at studying the trend/change in the water quality components over three years and predict the future trend. The study uses PCA to reduce the dimensions of the data and analyse it. The first step was to find the number of principle components to be formed. This was determined by the amount of variance that the components explain. The numbers of components retained are 4, 3, and 3 for the years 2017, 2018 and 2019 respectively.

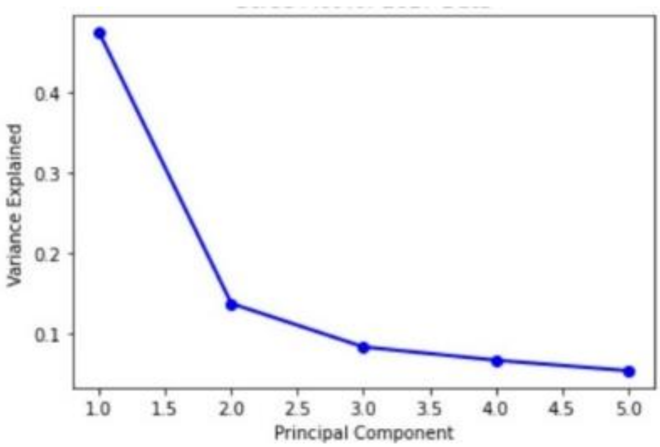


Fig. 2(a). Scree plot for 2017data

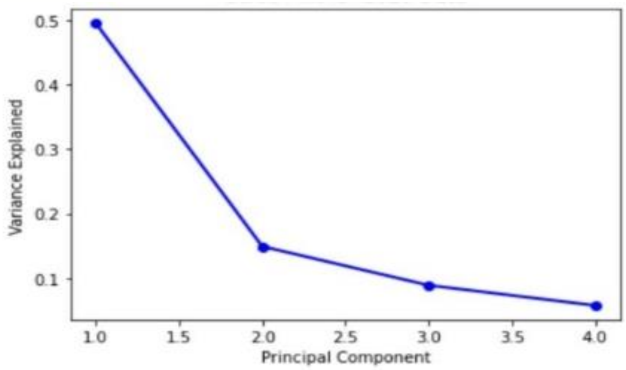


Fig. 2(b). Scree plot for 2018 data

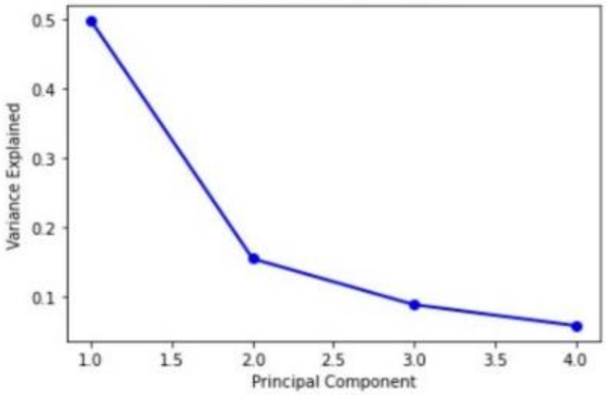


Fig. 2(c). Scree plot for 2019 data

The selection of the number of components is based on the scree plots shown in figures 2(a) (2b) (2c), which depicts the variability of the components for the years 2017, 2018 and 2019 respectively. The first component explains maximum variability compared to the next few components, which explain the moderate amount, and latter components show very little variability of the overall variance. The scree plot criterion looks for the "elbow" in the curve and selects all components just before the line flattens out.

Tables 3(a), 3(b) and 3(c) illustrate the calculated principal component loadings for each year.

Table 3(a). PCA results for 2017 data

Variable	PC1	PC2	PC3	PC4
TH	0.327	0.013	-0.014	-0.051
Ca	0.319	-0.002	-0.062	-0.074
Mg	0.327	0.034	0.020	-0.027
Na	0.124	-0.470	0.124	0.076
K	0.034	-0.354	-0.421	0.252
Fe	-0.014	-0.132	-0.590	-0.153
HCO_3	0.163	-0.279	0.369	-0.109
CO_3	0.026	-0.307	0.140	0.063
Cl	0.318	0.096	0.050	0.077
NO_3	0.168	-0.236	0.304	0.370
SO_4	0.204	-0.207	-0.367	-0.301
PO_4	0.285	0.045	-0.077	0.018
TDS	0.325	-0.077	-0.021	-0.025
EC	0.327	-0.089	-0.028	-0.030
F	0.224	0.311	0.017	-0.071
Cu	-0.039	-0.042	-0.214	0.728
pH	-0.000	-0.000	-0.000	0.000
Cr	0.225	0.314	0.001	0.162
pH	-0.042	-0.337	0.107	-0.272
Turbidity	0.265	0.167	-0.067	0.109

Table 3(b). Results of PCA for 2018 data

Variable	PC1	PC2	PC3
TH	0.360	0.428	-0.560
Ca	0.064	0.054	-0.177
Mg	0.048	0.071	-0.037
Na	0.017	-0.162	0.267
K	0.000	-0.002	-0.000
Fe	-0.000	-0.001	-0.001
HCO_3	0.027	-0.095	0.359
CO_3	0.000	-0.002	0.009
Cl	0.225	0.564	0.146
NO_3	0.016	0.016	0.263
SO_4	0.089	-0.625	-0.512
PO_4	0.001	0.001	-0.001
TDS	0.474	-0.129	0.128
EC	0.762	-0.218	0.211
F	0.000	0.003	-0.003

Cr	0.000	0.001	0.000
pH	-0.000	-0.001	0.001
Turbidity	0.001	0.002	-0.003

Table 3(c): PCA Results for 2019 data

Variable	PC1	PC2	PC2
TH	0.327	-0.017	-0.037
Ca	0.318	-0.001	-0.089
Mg	0.328	-0.031	0.007
Na	0.139	0.463	0.092
K	0.039	0.314	-0.404
Fe	-0.013	0.118	-0.599
HCO_3	0.136	0.336	0.365
CO_3	0.043	0.324	0.172
Cl	0.319	-0.084	0.057
NO_3	0.176	0.239	0.346
SO_4	0.208	0.195	-0.400
PO_4	0.287	-0.084	-0.058
TDS	0.327	0.084	-0.038
EC	0.328	0.082	-0.036
F	0.217	-0.322	0.002
Cr	0.225	-0.306	0.089
pH	-0.045	0.310	0.021
Turbidity	0.267	-0.195	0.007

3.1.1 Water quality trend in 2017

The results show that first component (PC1) is the indication of the measurement of salts in the water contributing to the variance of 47% of total 76% variance. TH **(**0.327**)**, Ca0.319), Mg (0.327), Cl (0.318), TDS (0.325), EC (0.328) are highly corelated components with loadings that are highlighted with bold in the table 2(a). Second component (PC2) accounts to 13% of variability in association with negative loading of sodium (-0.470) this component is the measure of decrease in sodium. Third component (PC3) accounts for 8% variance with strong negative loading of iron (-0.57). This component is the measure of decrease in the iron content in water. The fourth component (PC4) shows 6% variance in association of copper with strong positive loading of 0.728. Copper can get into the drinking water, as it passes through plumbing system. Plumbing systems with copper parts fewer than three years old usually would not have the time to build up this protective coating.

3.1.2 Water quality trend in 2018

The results show that the first component (PC1) is highly influenced by the electrical conductivity of water with a strong positive loading of 0.762 and moderately influenced by Total harness (0.36). This component accounts for 47% of total variance of 73%. Compared to 2017 trends, there is a marginal reduction in salts like TH, Ca and Mg. The second component (PC2) which contributes to 15% of variance is influenced by the strong negative loading of sulphate (-0.625) and positive loading of Cl (0.564). This component is an indication of decrease in sulphate. The third component (PC3) contributes to 8% of variance, continuing the negative influence with component of sulphate (-0.512) and also total hardness (-0.560). This component shows the measure of decrease in the total hardness of the water. The Table 2(b) also shows major change in the pH value compared to 2017 results. There is negligible impact of pH on the components.

3.1.3 Water quality trend in 2019

The first component (PC1) accounts to 50% variance of total variance of 73% and is associated with positive loading of salts like total hardness, calcium and magnesium, chloride), and electrical conductivity). It shows that amount of salts in 2017 has again recovered in 2019. The second component (PC2) accounts to 15% of variance associated with positive loading of 0.46 on sodium and moderate on bicarbonates (0.33). Third component (PC3) accounts to 8% of variance and is associated with negative loading of Iron in the water. This component is the measurement of decrease of iron in the water.

In second phase of the analysis, important water quality parameters were considered for study and also the first phase outcomes aided in the extraction of influential components. The variation of these components with respect to mean, standard deviation, minimum and maximum is studied as shown in the Table 4. This also helps in understanding the distribution of values across three years.

Table 4. Statistical Analysis of Critical Water Components

Variable	Mean			Standard Deviation			Minimum			Maximum		
	2017	2018	2019	2017	2018	2019	2017	2018	2019	2017	2018	2019
TH	757.96	751.28	781.86	691.06	681.47	700.7	88.0	92.0	98.0	3111.0	2982	3070
Ca	152.68	156.96	163.16	122.79	124.57	127.73	18.0	21.0	25.00	582.0	582	596.0
Mg	85.00	87.51	91.20	90.31	92.19	94.83	6.0	7.0	7.00	402.00	402.0	412.0
Fe	0.25	0.26	0.27	0.37	0.37	0.38	0.0	0.0	0.0	0.0	1.49	1.51
Cl	399.51	406.83	417.6	447.65	443.9	449.5	49.00	49	40.0	2142.00	2108.0	2120.0
NO_3	52.41	53.93	57.35	60.23	60.67	61.78	8.00	8.0	10.0	332.00	332.0	344.0
TDS	1153.0	1192.6	1237.5	866.86	877.51	902.32	210.0	230.0	248.0	3940.0	3940.0	4010
F	1.44	1.45	1.47	1.37	1.38	1.38	0.2	0.2	0.14	6.09	6.09	6.12
Cr	0.09	0.10	0.10	0.51	0.50	0.51	0.0	0.0	0.0	3.00	2.84	2.86
pH	7.22	7.22	7.22	0.72	0.73	0.73	5.1	5.1	5.12	8.24	8.25	8.30

The Figures 3(a), 3(b) and 3(c) helps in visualizing the trend of sample components (TH, Ca, Mg) over three years. The results show that there is an upward/rising trend in the water quality components.

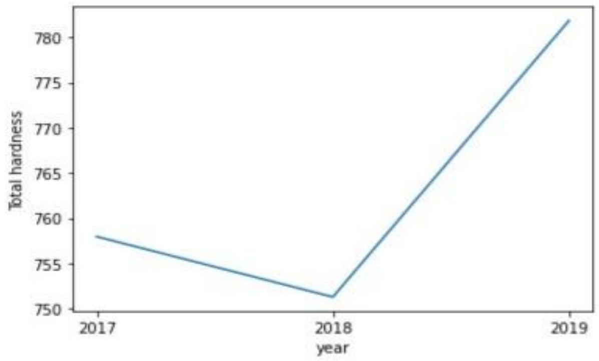


Fig. 3(a). Trend of total hardness through 2017-19

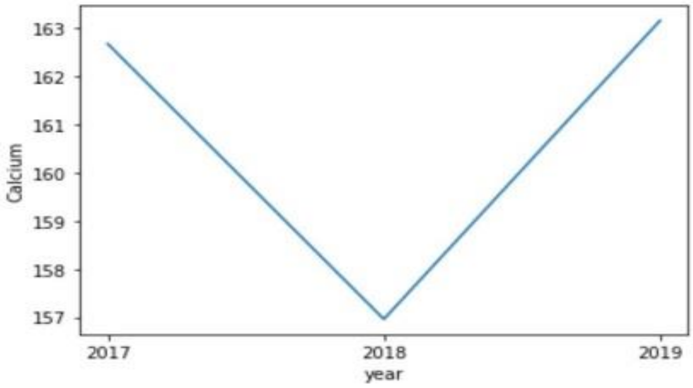


Fig. 3(b). Trend of calcium through 2017-19

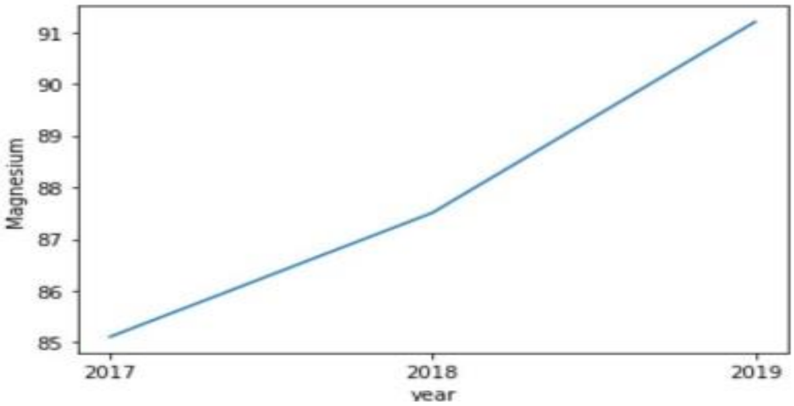


Fig. 3(c). Trend of magnesium through 2017-19

The third phase of the work deals with predicting the values of the critical components for the next few years based on the previous trend. Figures 4.1 to 4.5 show the future trend of the critical water components (Ca, NO_3 ,pH,Cl,Fe)

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f23d2b0ae90>
```

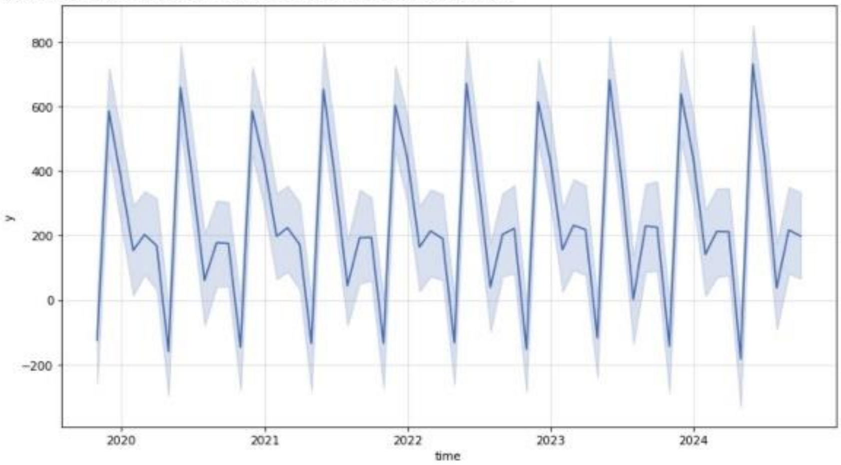


Fig. 4.1. Future trend of calcium through 2022-24

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa35d0542d0>
```

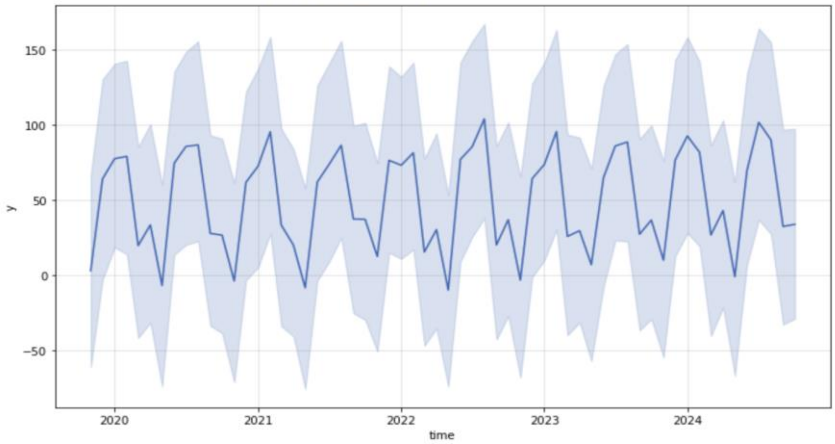


Fig. 4.2. Future trend of nitrate through 2022-24

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e52ef6b10>
```

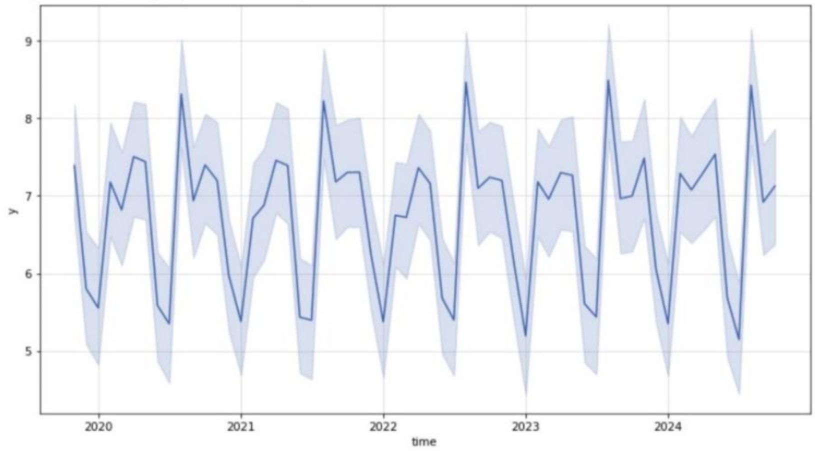
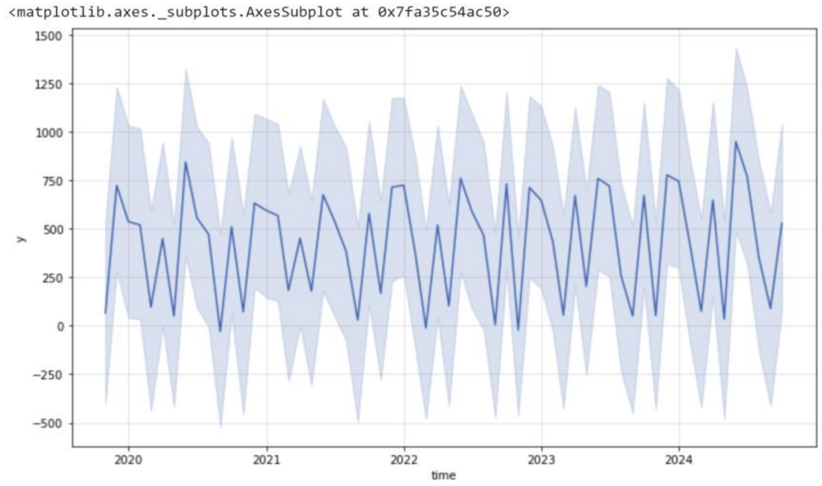
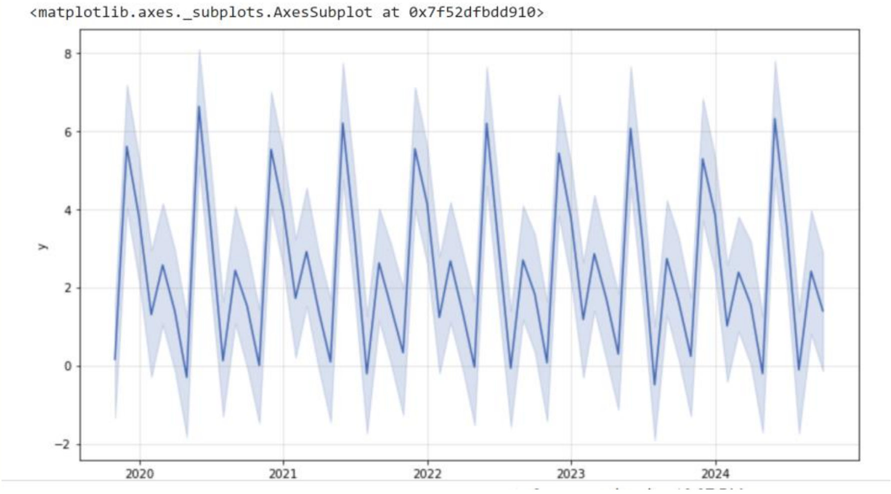


Fig. 4.3. Future trend of pH through 2022-24



```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa35c54ac50>
```

Fig. 4.4. Future trend of chloride through 2022-24



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f52dfbdd910>
```

Fig. 4.5. Future trend of iron through 2022-24

Conclusions

The physico-chemical analysis investigation carried out for the groundwaters of Peenya industrial area for a data set of 20 water quality parameters across 30 stations during the pre and post monsoon seasons of the years 2017, 2018, and 2019 revealed that 76.67% of the groundwater samples were not fit for drinking. Statistical analysis was carried out using Principal Component Analysis and time series analysis using python programming language. Since the study area is highly prone to industrial effluent discharges, the present work focused on the studying the changes in water quality parameters that happened over 3 years. The results revealed that there is an increasing tendency of the water quality parameter concentrations, though the variations are marginal. The study used time-series for the prediction of water quality components for the next few years, which would give an insight to the expected quality changes for the upcoming years and aid in undertaking appropriate measures to control/ reduce the groundwater contamination.

Acknowledgements

The authors are extremely grateful to the management and Principal of Cambridge

Institute of Technology for their wholehearted support and facilities extended during the course of this work.

References

[1] Aburub F, Hadi W. (2016) Predicting Groundwater Areas Using Data Mining Techniques: Groundwater in Jordan as Case Study. 2016;10.

[2] APHA (2002). Standard methods for the examination of water and wastewater (20th ed.). Washington D.C, New York, USA: American Public and Health Association.

[3] Assessing Groundwater Quality: A Global Perspective: Importance, Methods and Potential Data Sources. A report by the Friends of Groundwater in the World Water Quality Alliance. Information Document Annex for display at the 5th Session of the United Nations Environment Assembly. World Water Quality Alliance. 2021.

[4] Buitinck et al., (2013) API design for machine learning software: experiences from the scikit-learn project, 2013.

[5] Bureau of Indian Standards. In: BIS 10500; 2012.

[6] Christina Rajam Vijayakumar, Divya Priya Balasubramani, Hazi Mohammad Azamathulla. (2021) Assessment of groundwater quality and human health risk associated with chromium exposure in the industrial area of Ranipet, Tamil Nadu, India. Journal of Water, Sanitation and Hygiene for Development (2022) 12 (1): 58–67. ttps://doi.org/10.2166/washdev.2021.260.

[7] Deka, P. C. (2019) A Primer on Machine Learning Applications in Civil Engineering. 1st edn. CRC Press. Available at: https://www.perlego.com/book/1472674/a-primer-on-machine-learning-applications-in-civil-engineering-pdf (Accessed: 14 October 2022).

[8] Elhag M, Gitas I, Othman A. (2021) Time series analysis of remotely sensed water quality parameters in arid environments. Saudi Arabia; Environment Development and Sustainability. 2021. Available from: https://link.springer. com/article/10.1007/s10668-020-00626-z.

[9] Gasim Hayder, Isman Kurniawan, Hauwa Mohammed Mustafa. (2021) Implementation of Machine Learning Methods for Monitoring and Predicting Water Quality Parameters. 2021: 11(2): 9285 – 9295, https://doi.org/10.33263/BRIAC112.92859295.

[10] Giao T, Anh NK, Nhien PTH, et al. (2022) Groundwater Quality Assessment Using Groundwater Quality Index and Multivariate Statistical Methods and Human Health Risk Assessment in a Coastal Region of the Vietnamese Mekong Delta. Applied Environmental Research. 2022;44(2):68–85.

[11] Gun JVD. (2012) Groundwater and Global Change: Trends, Opportunities and Challenges. 2012. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000215496.

[12] Hamed, Reda MA, (2019) editors. Application of surface water quality classification models using principal components analysis and cluster analysis available at; 2019. Available from: https://ssrn.com/abstract=3364401orhttp://dx.doi.org/10.2139/ssrn.3364401.

[13] Hamid A, Bhat SA, Bhat SU, (2016) editors. Environmetric techniques in water quality assessment and monitoring; 2016. DOI 10.1007/s12665-015-5139-3.

[14] Hasan BMS, Abdulazeez AM, editors. (2021) A review of principal component analysis algorithm for dimensionality reduction..2021. Available from: https://doi.org/10.30880/ jscdm.2021.02.01.003.

[15] Hulya Boyacioglu, Hayal Boyacioglu, (2017) Application of environmetric methods to investigate control factors on water quality. DOI 10.1515/aep-2017-0026.

[16] Kofi Sarpong Adu-manu, Cristiano Tapparello, Wendi Heinzelman, Ferdinand Apietu

Katsriku and Jamal-deen Abdulai. (2017) Water quality monitoring using wireless sensor networks: current trends and future research directions, acm transactions on sensor networks, vol. 13, no. 1, article 4, publication date: January 2017.

[17] Lapworth DJ, Macdonald AM, Kebede S, et al. (2020) Drinking water quality from rural handpump-boreholes in Africa. Environmental Research Letters. 2020;15(6):64020–64020.

[18] Mahapatra SS, Sahu M, Patel RK. (2012) Prediction of Water Quality Using Principal Component Analysis. Water Qual Expo Health. 2012; 4:93–104.

[19] Mohammadpour A, Gharehchahi E, Badeenezhad A, Parseh I, Khaksefidi R, Golaki M, Dehbandi R., Azhdarpoor A., Derakhshan Z, Rodriguez-Chueca J., Giannakis S. (2022) Nitrate in Groundwater Resources of Hormozgan Province, Southern Iran: Concentration Estimation, Distribution and Probabilistic Health Risk Assessment Using Monte Carlo Simulation. Water **2022**, 14, 564. https://doi.org/10.3390/w14040564.

[20] Ravenscroft P, Mahmud ZH, Islam MS, et al. (2017) The public health significance of latrines discharging to groundwater used for drinking. Water Research. 2017; 124:192–201.

[21] Salema N, Husseinb S. (2019) Data dimensional reduction and principal components analysis. Procedia Computer Science. 2019; 163:292–299.

[22] Shankar B S, Usha Arcot. (2021) An index-based quality evaluation of groundwater—a case study of Whitefield area in Bangalore, India, International Journal of Environmental Analytical Chemistry, 2021. DOI: 10.1080/03067319.2021.2002311.

[23] Shankar BS, Balasubramanya N, Reddy MTM. (2008) Impact of industrialization on groundwater quality - a case study of Peenya industrial area. Environ Monit Assess. 2008; 142:263– 268.

[24] Sidra Mehtab, Jaydip Sen. (2020) A Time Series Analysis-Based Stock Price Prediction Using Machine Learning and Deep Learning Models. International Journal of Business Forecasting and Marketing Intelligence (IJBFMI), Vol 6, No 4, pp. 272 - 335, 2020. Inderscience Publishers, DOI: https://doi.org/10.1504/IJBFMI.2020.11569.

[25] Tizro A, Ghashghaie M. Pantazis Georgiou, Konstantinos Voudouris. (2014) Time series analysis of water quality parameters. Journal of Applied Research in Water and Wastewater. 2014; 1:43–52.