International Journal of Statistics and Applied Mathematics 2021; 6(6): 169-178

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2021; 6(6): 169-178

www.mathsjournal.com
Received: 11-09-2021
Accepted: 14-10-2021

Abukari Alhassan
Department of Statistics, University for Development Studies, Nyankpala Campus, Ghana

Katara Salifu
Department of Statistics, University for Development Studies, Nyankpala Campus, Ghana

Mahama Ishaque
Department of Economics, S. D. Dombo University of Business and Integrated Development Studies, Wa, Ghana

Amakye Ebenezer
Department of Statistics, University for Development Studies, Navrongo Campus, Ghana

Kaku Sagary Nokoe
Faculty of Economics and Business Administration, Catholic University College, Fiapre, Sunyani, Ghana

Corresponding Author:
Abukari Alhassan
Department of Statistics, University for Development Studies, Nyankpala Campus, Ghana

Groundwater quality modelling in upper Denkyira east municipality of the central region in Ghana

Abukari Alhassan, Katara Salifu, Mahama Ishaque, Amakye Ebenezer and Kaku Sagary Nokoe

Abstract
The aim of this study was to identify some common factors and develop a model for predicting Water Quality Index (WQI) in Upper Denkyira East Municipality of the Central region in Ghana. Factor Analysis with principal component method of extraction was used to identify the common factors and their factor scores were used as independent variables with WQI computed from the data as dependent variable to perform Multiple Linear Regression (MLR). Factor Analysis extracted hardness (Mn, Cd, Ca^{2+} and HCO_3^-), heavy metals (Pb, Cu, and As), salinity (Na^+, Cl^- and K^+), organic pollutant (NO_3^-, PO_4^{3-}, Zn and SO_4^{2-}) and chromium factor (Cr) as the common factors explaining almost 80% of the total variation in the data set. The results of the diagnostic check performed on the MLR model indicated that the model's accuracy was good for predicting WQI yielding an adjusted R^2 of 0.877. It was revealed that only the first three common factors (hardness, heavy metals, and salinity) were significant for predicting WQI. The results further showed that only salinity appeared to decrease WQI in the municipality. It is recommended that activities or factors contributing to higher concentration of hardness and heavy metals far above their recommended standard values in drinking water should be regulated in the municipality.

Keywords: Factor analysis, variance inflation factor, KMO and Bartlett's test, water quality index, groundwater quality

Introduction
Water is a standout among the most significant substances which plays an essential role in the everyday life of every single living being and impacts climatic changes and land shaping(Aleem *et al.*, 2018) [1]. Groundwater is a reliable wellspring of water supply since it is usually unpolluted because of controlled movement of contaminants in the soil profile. Nevertheless, shallow and permeable water table aquifers are most likely to be contaminated. Groundwater quality encompass the physical, chemical, and biological qualities of groundwater (Zhong *et al.*, 2006) [19]. Some of the numerous factors that can affect water quality include climate, geography, aquifer lithology, surface water recharge, saline water intrusion, human activities, etc.; occurring naturally or human occurring processes. (Bani, 2015; Huang *et al.*, 2008) [3, 10].
The Upper Denkyira East Municipal falls within the Central Region of Ghana. Ghana Statistical Service (GSS) reported in the 2010 population and housing census reported groundwater (borehole) as the major source of water for drinking and domestic use in the Municipality. However, as much as borehole water is the most widely used and serving as a source of water supply in the municipality, still have several factors affecting its usage. The major human activities in the Municipality are farming (crop farming) and mining (illegal mining) which can affect the groundwater quality and render it unsafe for drinking, domestic use etc. It is therefore paramount for researchers to examine the groundwater quality status in the municipality. A common approach employed to examine or determine the status of a source of water for safety used is WQI. WQI is a technique employed by many hydrologists in understanding the general water quality state and hence, has been relevant for assessing surface and groundwater quality everywhere in the world. (Minakshi and Dulal, 2016) [12]. However, groundwater quality is controlled by several water quality parameters which need to be sampled and analyse. Sampling and analyzing numerous groundwater quality parameters from many stations is not only difficult but also costly and time consuming.

Nevertheless, there are some latent factors (common factors) that have much influence on groundwater quality, which can be good representative of several groundwater quality parameters for determining the water quality status at a specific area. This study therefore seeks to apply Factor Analysis (FA) on several groundwater quality parameters sampled from different stations within Upper Denkyira East Municipality to identify the common factors that are most influential on groundwater quality and develop a Multiple Linear Regression (MLR) model based on the factor scores of the common factors for predicting WQI in the Municipality. Water quality studies that employed FA and MLR include Fahmi *et al.*, (2011) [7], Ewaid *et al.*, (2018) [6], Amiri and Nakane (2009) [2], Eslamian *et al.*, (2010) [5], Petersen *et al.*, (2001) [15] and Koklu *et al.*, (2010) [11] among others.

2. Materials and Methods

2.1 Study Area

The Upper Denkyira East Municipality forms part of the 22 Metropolitan, Municipal and District Assemblies (MMDAs) in Central Region of Ghana. It covers a total land space of 1,020 square kilometers, which is around 10% of the full land space of the Central Region and exists in Latitudes 5º. 30' and 6º 02' North of the Equator and Longitudes 1º W and 2º West of the Greenwich Meridian. The Municipality stocks normal limits with Adansi South in the north, Assin Central Municipal in the east and Twiffo Atti-Morkwa District in the west and Upper Denkyira West District in the north-west. (ghanadistricts.com). The region falls under a forest took apart level, developing to around 250m above ocean level. There are pockets of steep sided slopes substituting with level - base valleys. The significant stream inside the spot is the River Offin. Some of streams which are feeders of either waterways Offin and Pra move through the area. (mofa.gov.gh). The municipality has rich mineral stores particularly alluvial gold store along the valleys of River Offin and its feeders and gold stores inland (mofa.gov.gh).

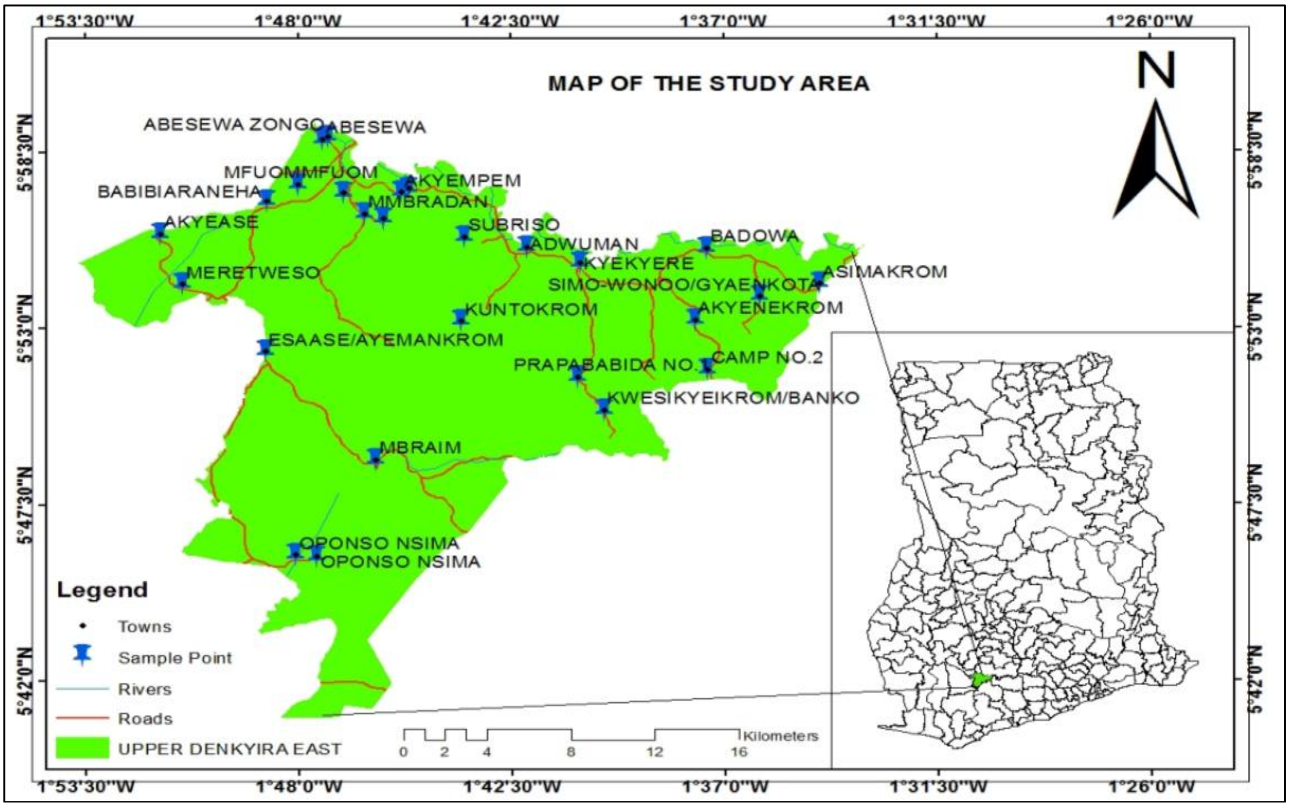

Fig 2.1: Study Area and Sampling Locations

2.2 Source of Data and Analysis

The data used in this study is a secondary data from records of Earth Science Department, University For Development Studies. It was collected from boreholes of 30 selected communities (sampled stations) within Upper Denkyira East Municipality for the year 2017. Each sample point (station) consist of 20 water quality parameters (Sodium (Na^+), Potassium (K^+), Nitrate (NO_3^-), Phosphate (PO_4^{3-}), Sulphate (SO_4^{2-}), Carbonate (CO_3^{2-}), Bicarbonate (HCO_3^-), Chloride (Cl^-), Calcium (Ca^{2+}), Magnesium (Mg^{2+}), Manganese (Mn), Iron (Fe), Zinc (Zn), Copper (Cu), Cadmium (Cd), Chromium (Cr), Nickel (Ni), Cobalt (Co), Lead (Pb), and Arsenic (As)) with their measured concentrations in milligram per liter (mg/l). Statistical analysis such as Factor Analysis and MLR were performed using IBM SPSS Statistics (version 20) statistical software, and Microsoft office Excel 2013 was used to compute the WQI of all the sampled stations. Nickel (Ni) and cobalt (Co) were removed from further analysis because they recorded a constant value for all the sampled stations, and CO_3^{2-} was also eliminated due to its perfect correlation with HCO_3^-.

2.3 Factor Analysis (FA)

Factor Analysis (FA) is a data reduction techniques that seeks to account for the covariance or correlation among observed variables. It is used to derive small number of factors called common factors that describe most of the variance that is observed in a larger number of variables. In FA, the observed variables are expressed as linear combination of the common factors. Suppose we have p variables $y_1, y_2 \ldots y_p$ with mean vector μ and variance covariance ($\sum$). We assume that μ = 0, $\sum$ is of full rank, and Var(f_i) = 0,. The factor equation is given as,

$$y_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \dots + \lambda_{jm} f_m + \varepsilon_j \qquad 2.1$$

$j = 1, 2, \dots, p$ and $i = 1, 2, \dots, m$, $m < p$, but in this study p is 17

Where:
m is the number of common factors; ε_j is the error term; f_i are the common factors, and λ_{ji} are the factor loadings which shows how each y_j individually depends on the f's.
Taken variance of equation 2.1 becomes:

$$Var(y_j) = \sum_{k=1}^{m} \lambda_{jk}^2 + \psi \qquad 2.2$$

Where:

$\sum \lambda_{jk}^2$ is the proportion of variance in the variable that is explained by the common factors (communality).
ψ is the unique variance for a particular variable. Equation 2.1 can be written in matrix notation as:

$$y = \Lambda f + \varepsilon \qquad 2.3$$

Where

y = $(y_1, y_2, \dots, y_p)'$, $f = (f_1, f_2, \dots, f_m)'$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots \varepsilon_p)'$ and

$$\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{pmatrix}$$

The principal components extraction technique was used to extract the factors. In this technique, the factors are related to the first m components where the j^{th} loadings are the scaled coefficients of the j^{th} principal components. The factors are interpreted the same way in principal components analysis if the components are not rotated. The principal component factor analysis of the sample covariance matrix S (or correlation matrix R) is specified in terms of its eigenvalue - eigenvector pairs (λ_i, e_i), $i = 1, \dots, p$ and $\lambda_1 \le \lambda_2 \le \dots \le \lambda_p$. Let $m < p$ be the number of common factors. The estimated factor loadings matrix is a $p \times m$ matrix, L, whose i^{th} column is, $i = 1, \dots, m$. (Minitab, 2016).
A varimax orthogonal rotation was employed to rotate the factor loadings. In this rotation approach, the rotated loadings maintain the residual matrix, the communalities and the specific variances, but maximizes variance of squared loadings inside factors which simplifies the columns of the loading matrix. Rotation places the axes near as many points as possible and links each group of variables with a factor. This helps to minimize high cross -loadings of variables on the factors and make interpretation easier and theoretically meaningful. (Rencher, 2002) [16].
The rotated loading matrix is given as.

$$\hat{\Lambda}^* = \hat{\Lambda} T \qquad 2.4$$

Where T is the orthogonal matrix, and $\hat{\Lambda}$ is the estimate of Λ in equation 2.3

The Factor scores ($\hat{f}_i$) are defined as estimates of the common factors. They show the contribution of each of the common factors on each of the sampled stations. They can be used as inputs for further analysis. In this study, the factor scores were used as independent variables for MLR analysis. The regression approach was used to compute the factor scores and it is given as:

$$\hat{F} = Y_s R^{-1} \hat{\Lambda}^* \qquad 2.5$$

Where

$\hat{F} = \left(\hat{f}_1, \hat{f}_2, \dots, \hat{f}_n\right)'$, n is the number of observations. (Rencher, 2002) [16]. In this study our n (sampled stations) is 30. R is the correlation matrix, $\hat{\Lambda}^*$ is defined in equation 2.4 and

Y_S is the observed matrix of standardized variables, $\frac{(y_{ij} - \bar{y}_j)}{s_j}$

Assumptions
Statistical assumptions in Factor Analysis includes:
1. Variables are multivariate normal
2. Linear relations between variables
3. Significant correlation among variables

However, principal component method of extraction which was used in this study does not make strong assumption on the distribution of the variables.

Kaiser-Meyer-Olkin and Bartlett's Test
The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is used to check the suitability of factor analysis for the data as well as the partial correlations among the variables. KMO values greater than or equal to 0.5 (average value) means that factor analysis is suitable for the data. (Wikiversity, 2019) [18].
The Bartlett test shows the strength of the association between variables and indicates whether or not the data is appropriate for Factor Analysis. The null hypothesis is stated as: the correlation matrix is an identity matrix. A p – value less that alpha value of 0.05 means that the null hypothesis is false. We therefore reject H_o and conclude that there is significant correlation among the variables and hence the Factor Analysis is suitable for the data.

2.4 Multiple Linear Regression
A Multiple Linear Regression model is a regression model with one dependent variable and two or more independent variables. It is used to explain the relationship between the dependent variable with the independent variables as well as predict the dependent variable using the independent variables. It is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \qquad 2.6$$

Where

β_0 = Regression Constant

β_j = Regression Coefficient for variable $\chi; j = 1, 2, \ldots, k$
k = Number of Independent Variables
ε = Residual
Estimated Multiple Linear Regression model is given by:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \quad 2.7$$

In this study, WQI computed from the data was used as the dependent variable and the factor scores of the 5 extracted factors were used as the independent variables. The independent variables were defined as:
f_1 = hardness, f_2 = heavy metals, f_3 = salinity, f_4 = organic pollutants, and f_5 = chromium factor.
The regression equation for this study can therefore be written as:

$$WQI = b_0 + b_1 f_1 + b_2 f_2 + b_3 f_3 + b_4 f_4 + b_5 f_5 \quad 2.8$$

Where b_j are the coefficients of the factors; $j = 1, 2, \ldots, 5$, and is interpreted as the predicted change in the WQI when the independent variable is increased by 1 unit given that all the other independent variables in the model are constant.

Assumptions
Multiple Linear Regression comes with the assumptions below:

1. The model error terms are normally distributed.
The Shapiro–Wilks (W) test was used to test for normality of the error terms. The null hypothesis is stated as:
H_0: Residuals are normally distributed.
At alpha level of 0.05, a *p*-value greater than 0.05 indicates that the null hypothesis is true.

2. The Residuals have a constant variance (Homoscedasticity)
This assumption was tested using Breusch – Pagan's (BP) Test. BP test was performed to identify problem of heteroscedasticity. The null hypothesis is stated as:
H_0: Residuals have constant variance
If significant – value is greater than alpha level of 0.05 we fail to reject the null hypothesis and conclude that the residuals have constant variance.

3. No Multicollinearity among the independent variables
Multicollinearity was tested using the Variance Inflation Factor (VIF). The VIF measures how much the variance of an estimated regression coefficient increases if the independent variables are correlated. A VIF value less than 10 for a given independent variable indicates that this independent variable is not significantly correlated with the remaining independent variables in the model. A value greater than 10 is an indication of multicollinearity. However, a value of 1.0 for a particular independent variable indicates that the correlation between this variable and the remaining independent variables is zero (0). VIF is given by the formula below:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \quad 2.9$$

Where:
R_j^2 = Coefficient of determination when the j^{th} independent variable is regressed against the remaining $k–1$ independent variables.

1. Linear relationship between the dependent and the independent variables.
2. The error terms are independent of each other.

2.5 Water Quality Index
Water quality index (WQI) is an essential mathematical tool which provides a single value that is used to ascertain the overall water quality status at a specific locality based on several water quality parameters. It expresses the water quality in simple terms such as excellent, good, poor etc. There are various methods for computing water quality index, but in this study we considered the Weighted Arithmetic Water Quality Index (WAWQI) approach. The equation according to Shweta *et al.*, (2013) is given as:

$$WQI = \frac{\sum Q_i W_i}{\sum W_i} \quad 2.10$$

Where Q_i is the quality rating scale and is given as:

$$Q_i = \left[\frac{(V_i - V_0)}{(S_i - V_0)}\right] \times 100$$

Where:
V_i = Estimated concentration of i^{th} parameter in the analysed water
V_0 = Ideal value of the parameter in pure water, (V_0 = 0, except PH = 7.0 and DO =14.6 mg/l)
S_i= Recommended standard value of i^{th} parameter. In this study we used Bhushan Mahajan WHO recommended standard value of the parameters.
W_i is the unit weight for each water quality parameter and is given by:

$$W_i = \frac{K}{S_i}$$

Where K is proportionality constant and can be calculated as:

$$K = \frac{1}{\sum \frac{1}{S_i}}$$

Therefore, from equation 2.10 we can say that the higher the concentration of a parameter above its recommended standard value the higher the WQI (poor water quality), and the lower the concentration of the parameter below its recommended standard value the lower the WQI (good water quality).

Table 2.1: Water Quality Rating

WQI Value	Rating of Water Quality
0 - 25	Excellent
26 - 50	Good
51 - 75	Poor
76 - 100	Very Poor
> 100	Unsuitable for Drinking

Source: Shweta *et al.*, (2013)

3. Results and Discussion

3.1 Descriptive Statistics

Table 3.1 shows the descriptive statistics of the various water quality parameters. At this stage of preliminary analysis, the results of the table indicate that Na^+, K^+, NO_3^-, SO_4^{2-}, CO_3^{2-}, HCO_3^-, Cl^-, Ca^{2+}, and Mg^{2+} are the most dominant parameters with mean concentration of 52.0333 mg/L, 11.2733 mg/L, 16.8748 mg/L, 6.2261 mg/L, 175.0825 mg/L, 178.0005 mg/L, 16.2616 mg/L, 17.4400 mg/L, and 10.6326 mg/L respectively, which is an indication that they are likely to come from a common source. However, the results of the standard deviation and the range indicate that variation among the measured values of these parameters at different stations is too high and variation range is wider than that of PO_4^{2-}, Fe, Zn, Cu, Mn, Cd, Cr, Ni, Co, Pb which are less dominant with mean concentration of 0.5924 mg/L, 0.0709 mg/L, 0.0767 mg/L, 0.0027 mg/L 0.0443 mg/L, 0.0029 mg/L 0.0102 mg/L, 0.0005 mg/L, 0.0025 mg/L, 0.0040 mg/L, and 0.0117 mg/L respectively.

Table 3.1: Descriptive Statistics of Water Quality Parameters

Parameter (mg/L)	N	Range	Minimum	Maximum	Mean	Std. Deviation
Sodium (Na^+)	30	96.5000	16.0000	112.5000	52.0333	18.1640
Potassium (K^+)	30	14.8000	6.7000	21.5000	11.2733	3.7523
Nitrate (NO_3^-)	30	449.9589	0.1031	450.0620	16.8748	81.8259
Phosphate (PO_4^{3-})	30	4.0784	0.0076	4.0860	0.5924	0.8375
Sulphate (SO_4^{2-})	30	13.0437	0.0520	13.0957	6.2261	3.1967
Carbonate (CO_3^{2-})	30	508.4587	7.1952	515.6539	175.0825	118.6675
Bicarbonate (HCO_3^-)	30	516.9330	7.3151	524.2481	178.0005	120.6453
Chloride (Cl^-)	30	93.9709	0.0000	93.9709	16.2616	18.6359
Calcium (Ca^{2+})	30	62.4000	1.6000	64.0000	17.4400	14.3684
Magnesium (Mg^{2+})	30	23.3204	0.0012	23.3216	10.6326	6.1316
Iron (Fe)	30	1.4315	0.0025	1.4340	0.0709	0.2686
Zinc (Zn)	30	0.5050	0.0000	0.5050	0.0767	0.1108
Copper (Cu)	30	0.0285	0.0000	0.0285	0.0027	0.0050
Manganese (Mn)	30	0.4725	0.0000	0.4725	0.0443	0.0953
Cadmium (Cd)	30	0.0560	0.0000	0.0560	0.0029	0.0101
Chromium (Cr)	30	0.1000	0.0000	0.1000	0.0102	0.0184
Nickel (Ni)	30	0.0000	0.0005	0.0005	0.0005	0.0000
Cobalt (Co)	30	0.0000	0.0025	0.0025	0.0025	0.0000
Lead (Pb)	30	0.0630	0.0005	0.0635	0.0040	0.0121
Arsenic (As)	30	0.1150	0.0005	0.1155	0.0117	0.0265

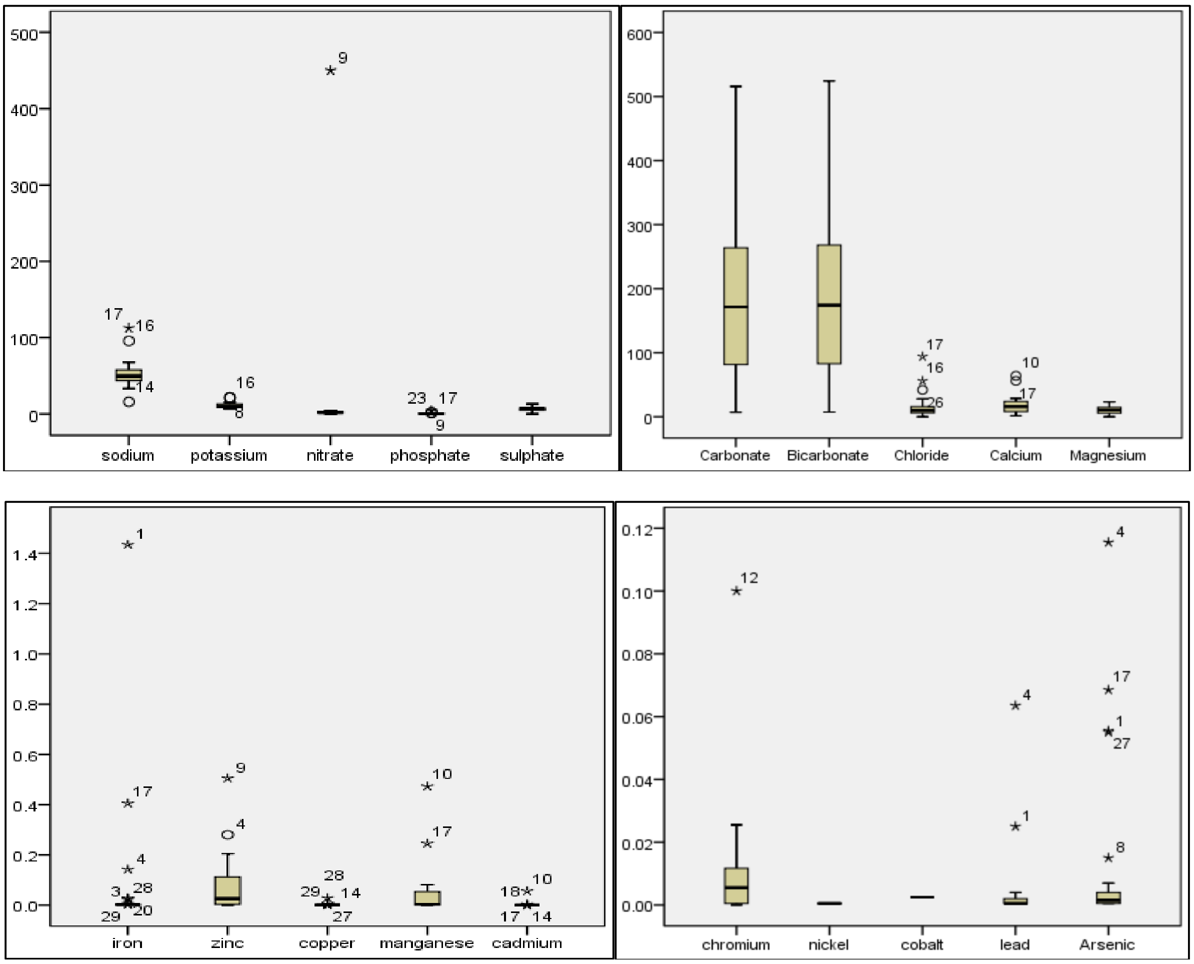

Fig 3.1: Boxplots of the Water Quality Parameters

3.2 Factor Analysis

3.2.1 Test of Suitability of Data for Factor Analysis

Table 3.2 shows the results of the KMO and Bartlett's Tests. From the table 3.2, which display a KMO value of 0.507 and the *p* – value of Bartlett's test of sphericity which tests the null hypothesis that the correlation matrix is an identity matrix is 0.000 which is less than alpha value of 0.05. Therefore the results of these tests indicate that the data is suitable for factor analysis

Table 3.2: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.507**
Bartlett's Test of Sphericity	Approx. Chi-Square	499.638
	df	136
	Sig.	0.000

3.2.2 Communalities

The communalities of the water quality parameters are shown in Table 3.3. The row labelled Extraction indicate the proportion of the variable's variance that have been accounted for by the common factors. Higher values above 0.5 indicate that the extracted components (factors) represents the variables very well. From the table only K^+ and Fe recorded weak communality.

Table 3.3: Communalities

Extraction	0.89	0.48	0.95	0.83	0.57	0.87	0.79	0.96	0.55	0.47	0.86	0.87	0.93	0.91	0.77	0.96	0.89
Initial	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Parameter	Na^+	K^+	NO_3^-	PO_4^{3-}	SO_4^{2-}	HCO_3^-	Cl^-	Ca^{2+}	Mg^{2+}	Fe	Zn	Cu	Mn	Cd	Cr	Pb	As

3.2.3 Number of Factors Extracted and Total Variance Explained

Using eigenvalue value greater than one rule and the scree plot, only 5 components (factors) were extracted as indicated in Table 3.4 and the scree plot (Figure 3.2) where components with eigenvalue greater than one are extracted as the common factors. The variance explained by the initial solution, extracted components, and rotated components are shown in Table 3.4. The second column of the Table indicates the extracted components; the first factor (F1) explained 29.542% of the total variance. The second factor (F2) accounted for 16.651%, the third factor (F3) explained 15.768%, the fourth factor (F4) described 10.942%; and the fifth factor (F5) accounted for 6.734% of the total variance. In all, the five factors acounted for approximately 80% of the total variance. This indicates that the dimension of the data set has been reduced by these five components (factors) with only 20% loss of information.

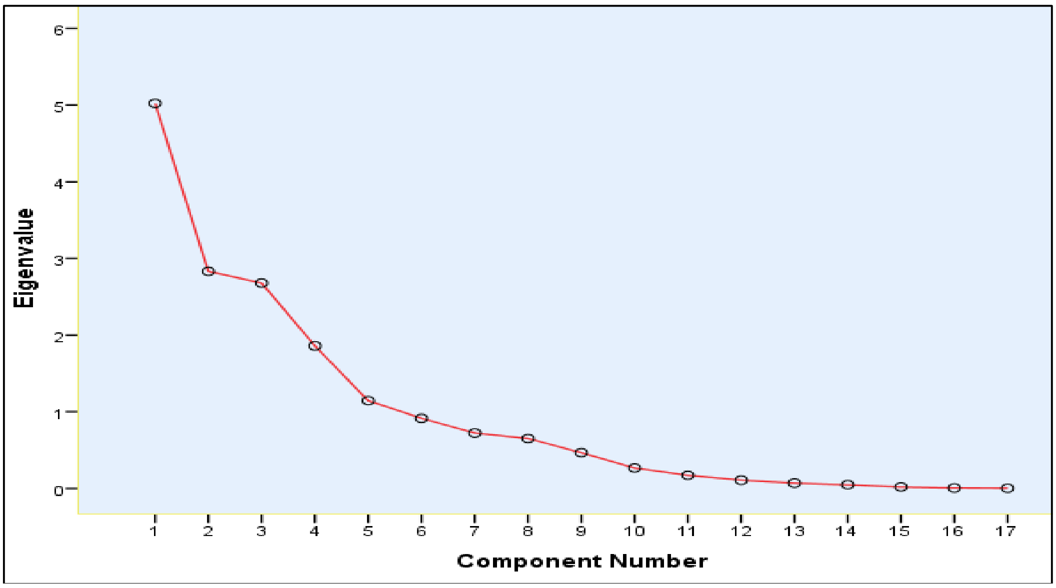

Fig 3.2: Scree plot

Table 3.4: Total Variance Explained

Component	Initial Eigenvalues			Extracted Sums of Squared Loadings			Rotated Sums of Squared Loadings		
	Total	Percent of Variance	Cumulative Percent	Total	Percent of Variance	Cumulative Percent	Total	Percent of Variance	Cumulative Percent
1	5.022	29.542	29.542	5.022	29.542	29.542	3.322	19.540	19.540
2	2.831	16.651	46.193	2.831	16.651	46.193	3.108	18.281	37.821
3	2.681	15.768	61.961	2.681	15.768	61.961	2.958	17.398	55.219
4	1.860	10.942	72.904	1.860	10.942	72.904	2.908	17.109	72.328
5	1.145	6.734	79.638	1.145	6.734	79.638	1.243	7.310	79.638
6	0.916	5.389	85.027						
7	0.724	4.258	89.285						

8	0.651	3.832	93.117						
9	0.468	2.755	95.872						
10	0.267	1.572	97.444						
11	0.171	1.008	98.451						
12	0.109	0.639	99.090						
13	0.071	0.419	99.509						
14	0.050	0.292	99.801						
15	0.021	0.124	99.925						
16	0.008	0.048	99.973						
17	0.005	0.027	100.000						

3.2.3 Rotated Component Matrix

Table 3.5 presents the rotated components matrix for the five factors after Varimax (orthogonal) rotation with Kaiser Normalization. Using factor loadings above 0.75 for strong loadings and 0.75 - 0.50 for moderate loadings (because for strong correlation r > 0.75 and for moderate correlation, r ≤ 0.50), it is shown in Table 3.5 that the factor 1 (F1), which explained 29.542% of the total variance, has strong positive loadings on HCO_3^-, Ca^{2+}, Mn and Cd. Thus Factor 1 is a mixture of ions and heavy metals and can be termed as hardness of water.

Factor 2 (F2), which accounts for 16.651% of the total variance had strong positive loadings on Cu, Pb and As. This factor is a mixture of copper (Cu), lead (Pb), and Arsenic (As), and can be labeled as heavy metals.

Factor 3 (F3), which accounted for 15.768% of the total variance recorded strong positive loadings on Na^+ and Cl^-, and moderate loading on K^+. Factor 3 (F3) can be named salinity.

The factor 4 (F4), which contributes 10.942% of the total variation loaded strongly on NO_3^-, PO_4^{3-}, and Zn and moderately on SO_4^{2-}. Factor 4 (F4) is known as organic pollutants.

The factor 5 (F5), which explains 6.734% of the total variance has strong negative loading (r < - 75) on Cr. This factor can also be labeled chromium factor.

Table 3.5: Rotated Component Matrix

Parameter		Na^+	K^+	NO_3^-	PO_4^{3-}	SO_4^{2-}	HCO_3^-	Cl^-	Ca^{2+}	Mg^{2+}	Fe	Zn	Cu	Mn	Cd	Cr	Pb	As
Component	1	0.28	0.04	-0.11	-0.11	0.13	0.82	0.04	0.85	0.22	0.06	0.22	0.00	0.93	0.90	-0.01	0.00	0.12
	2	0.02	0.09	-0.11	-0.14	0.20	0.32	-0.09	0.17	0.25	0.38	0.32	0.92	-0.08	-0.12	0.00	0.97	0.88
	3	0.90	0.51	-0.09	0.26	0.43	0.29	0.87	0.45	0.45	0.40	-0.04	-0.13	0.20	-0.26	0.14	-0.04	0.30
	4	0.05	0.12	0.96	0.85	0.57	-0.03	0.12	0.08	0.48	0.04	0.82	0.04	0.02	0.03	-0.06	0.05	0.04
	5	-0.03	0.44	0.03	-0.06	0.12	-0.11	-0.12	-0.05	-0.03	0.41	0.20	-0.08	0.10	0.09	-0.87	0.10	0.11

The visual representation of the rotated component matrix is presented in Figure 3.3, where all variables in each component that loaded highly (from moderate correlation upwards) are brought together in order of the degree of correlation with each other.

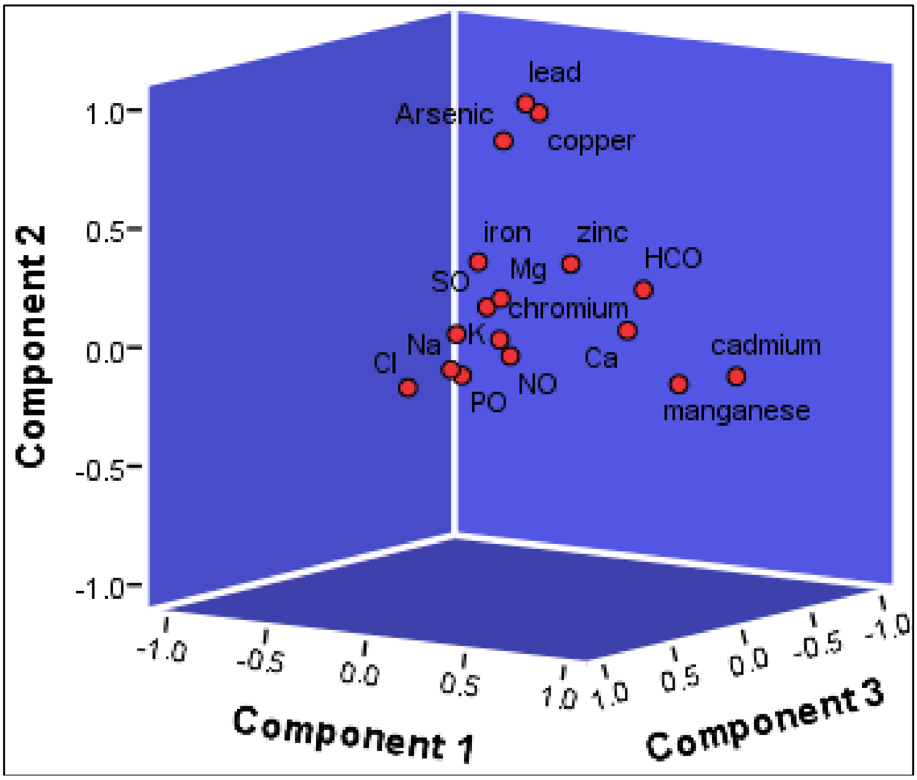

Fig 3.3: Component Plot in Rotated Space

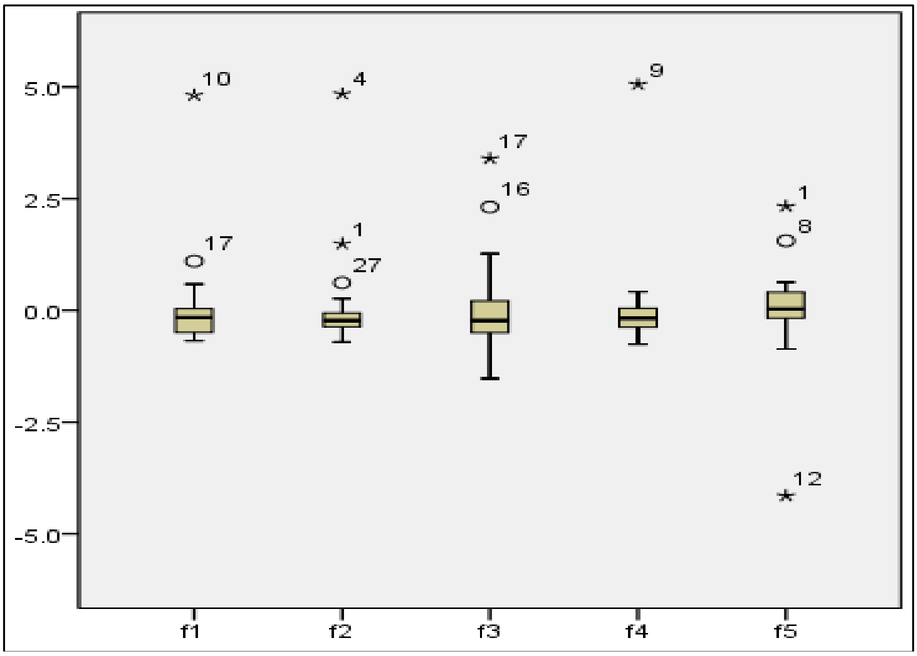

Fig 3.4: Boxplot of factor scores of the common factors

3.3 Multiple Linear Regression Analysis

3.3.1 Normality test

The Shapiro-Wilk test of normality is shown in Table 3.6. The *p*–value (significant value) of 0.697 which is greater than alpha value 0.05 indicating that the null hypothesis that the residuals are normally distributed is true. We therefore fail to reject the null hypothesis and conclude that the residuals are normally distributed.

Table 3.6: Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Unstandardized Residual	0.114	30	0.200*	0.975	30	0.697

3.3.2 Test of Homoscedasticity

Test of constant variance of the residuals is shown in Table 3.7. In the table, the Breusch-Pagan (BP) test of homoscedasticity which tests the null hypothesis that there is no heteroscedasticity gave a p-value of 0.443 which is greater than alpha level of 0.05 suggesting that the null hypothesis is true. We therefore conclude that the residuals have constant variance.

Table 3.7: Test of Homoscedasticity

Breusch-Pagan test statistics and significant values		
BP	LM	Sig
	4.785	0 .443

3.3.3 Strength and Overall Significance of the Regression Model

The fitness and overall significance of the model are summarized in Table 3.8 (Model Summary) and Table 3.9 (ANOVA) respectively. The model summary give the multiple correlation coefficient (*R*), R^2 and adjusted R^2. The multiple correlation coefficient $R = 0.948$ means that there is a very strong correlation between observed WQI and those predicted by the model. The adjusted $R^2 = 0.877$ indicates that approximately 88% of the total variance in the WQI have been explained by the common factors (hardness, heavy metals, salinity, organic pollutants, and chromium factor). The results of the ANOVA ($F(5, 24) = 42.433$, $p = 0.000 < 0.05$) suggests that the overall model is significant. Thus at least one of hardness, heavy metals, salinity, organic pollutants, and chromium factor is related to WQI

Table 3.8: Model Summary

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	0.948[a]	0.898	0.877	70.479536149

Table 3.9: Anova

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	1053912.825	5	210782.565	42.433	0.000[b]
Residual	119216.760	24	4967.365		
Total	1173129.585	29			

3.3.4 Coefficients Table

The output depicted in table 3.10 gives estimates of the Multiple Linear Regression model coefficients, standard errors of the estimates, *t*-tests that a coefficient takes the value zero (thus p – value > 0.05), and collinearity statistics. The Collinearity Statistics give the values of the tolerance and the variance inflation factor (VIF) of each of the independent variables. The value of the variance inflation factor, VIF =1.000 < 10 suggests that there is absence of multicollinearity.

The estimated coefficients of the regression model are given under the column of “Unstandardized Coefficients B”; these give, for each of the predictor variables, the predicted change in the WQI when the independent variable is increased by 1 unit given that all the other independent variables in the model are constant. From the coefficient output, hardness, f_1 with coefficient, B = 181.093 and significant value = 0.000 is less than 0.05 which is significant; its large positive coefficient indicates that, generally, hardness of the groundwater in the study area significantly increases water quality index (WQI). This is an indication that, largely, the concentration of hardness of water in the municipality is above its recommended standard value in water for drinking and other purposes.

Also, the effect of heavy metals, f_2 (B = 36.822, sig. value = 0.010 < 0.05) is significant and its coefficient is positive signifying that in general, the heavy metals in the municipality increase water quality index, thus the heavy metals decrease the groundwater quality for drinking and other purposes in the municipality. It can be attributed to higher concentration of the heavy metals above their recommended standard values in the municipality.

Again, the impact of salinity, f_3 (B = -33.991, sig. value = 0.016 < 0.05) on water quality index is significant, but its coefficient is highly negative suggesting that among the factors salinity appears to decrease the water quality index in the area. This means that the concentration of salinity in the groundwater is much lower than its recommended standard value in the municipality.

However, the effects of organic pollutants, f_4 (B = 21.911, significant. value = 0.107 > 0.05) and chromium factor, f_5 (B = 23.579, significant value = 0.084 > 0.05) are not significant in predicting water quality index (WQI), thus, their significant values are greater than 5% level of significance. Since the correlation between the independent variables are all zero (VIF = 1.0) the removal of the insignificant parameters will not change the regression coefficients of the significant parameters. The model is therefore given as:

$$WQI = 94.390 + 181.093f_1 + 36.822f_2 - 33.991f_3$$

Table 3.10: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			**Tolerance**	**VIF**
1	(Constant)	94.390	12.868	-	7.335	0.000	-	-
	f_1	181.093	13.088	0.900	13.837	0.000	1.000	1.000
	f_2	36.822	13.088	0.183	2.813	0.010	1.000	1.000
	f_3	-33.991	13.088	-0.169	-2.597	0.016	1.000	1.000
	f_4	21.911	13.088	0.109	1.674	0.107	1.000	1.000
	f_5	23.579	13.088	0.117	1.802	0.084	1.000	1.000

4.0 Conclusion

Twenty (20) water quality parameters from boreholes of 30 sampled stations within Upper Denkyira East Municipal in the Central region of Ghana were used in this study. Factor analysis identified hardness (Mn, Cd, Ca^{2+} and HCO_3^-), heavy metals (Pb, Cu, and As), salinity (Na^+, Cl^- and K^+), organic pollutants (NO_3^-, PO_4^{3-}, Zn and SO_4^{2-}) and chromium factor (Cr) as the common factors influencing groundwater quality in the municipality. These factors explained almost 80% of the total variation in the data set. The results of the diagnostic check performed on the multiple linear regression model indicated that the model has good accuracy for predicting water quality index with multiple correlation coefficient, R = 0.948 and adjusted R^2 = 0.877. The t-test for significance of individual parameters showed that only the first three common factors (hardness, heavy metals, and salinity) were significant for predicting water quality index. The results exposed that only salinity appears to decrease WQI (good water quality) which is an indication that the concentration of salinity is below its recommended standard value in the municipality. Hardness and heavy metals on the other hand increase WQI (poor water quality) and were attributed to higher concentration of these factors far above their recommended standard values in water, which could be due to the geology of the area, waste discharges from the mining industries and agricultural activities. It is recommended that activities or factors contributing to higher concentration of hardness and heavy metals far above their recommend standard values in drinking water should be regulated in the municipality. Also, it is recommended that researchers who wish to carry out research on water quality in the study area should take into consideration the significant common factors identified in this study.

References

1. Aleem M, Shun CJ, Li C, Aslam AM, Yang W, Nawaz MI, *et al*. Evaluation of groundwater quality in the vicinity of Khurrianwala industrial zone, Pakistan. Water (Switzerland). 2018;10(10):1-21. https://doi.org/10.3390/w10101321
2. Amiri BJ, Nakane K. Modeling the linkage between river water quality and landscape metrics in the Chugoku district of Japan. Water Resources Management. 2009;23(5):931-956. https://doi.org/10.1007/s11269-008-9307-z
3. Bani E. Investigating Anthropogenic Impact on Groundwater Quality in the Ga East. (10440206). 2015.
4. Bhushan M. Drinking Water Quality Standards (WHO Guidelines). Accessed from Drinking Water Quality Standards (WHO Guidelines) (civiconcepts.com)
5. Eslamian S, Ghasemizadeh M, Biabanaki M, Talebizadeh M. A Principal Component Regression Method for Estimating Low Flow Index. Water Resources Management. 2010;24(11):2553-2566. https://doi.org/10.1007/s11269-009-9567-2
6. Ewaid SH, Abed SA, Kadhum SA. Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. Environmental Technology & Innovation. 2018;11:390-398. https://doi.org/10.1016/J.ETI.2018.06.013
7. Fahmi M, Nasir M, Samsudin MS, Mohamad I, Roshide M, Awaluddin A. River Water Quality Modeling Using Combined Principle Component Analysis (PCA) and Multiple Linear Regressions (MLR): A Case Study at Klang River, Malaysia Department of Environmental Sciences, Faculty of Environmental Studies, Department of Enviro. 2002-11;14:73-82.
8. Ghana Statistical Service. District Analytical Report: Upper Denkyrira East Municipality. 2010 Population and Housing Census 2014, 74. Retrieved from www.statsghana.gov.gh/docfiles/2010.../UPPER DENKYIRA EAST.pdf%0A%0A
9. ghanadistricts.com. (n.d.). Ghana Districts: A repository of all Local Assemblies in Ghana. Retrieved June 2, 2019, from http://www.ghanadistricts.com/Home/District/76
10. Huang ML, Zhou SL, Sun B, Zhao QG. Heavy metals in wheat grain: Assessment of potential health risk for inhabitants in Kunshan, China. Science of the Total

Environment. 2008, 54-61.
11. Koklu R, Sengorur B, Topal B. Water quality assessment using multivariate statistical methods-a case study: Melen river system (Turkey). Water Resources Management. 2010;24(5):959-978. https://doi.org/10.1007/s11269-009-9481-7
12. Minakshi B, Dulal CG. Water Quality assessment in terms of Water Quality Index (WQI): case study of the Kolong, river, Assau, India. Apply Water Science. 2016-17;7:312-3135. https://doi.org/doi:10. 1007/s 13201 - 016 - 0451 - y
13. minitab.com. (n.d.). Methods and formulas for Factor Analysis. Retrieved May 31, 2019, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/factor-analysis/methods-and-formulas/methods-and-formulas/#fnsrc_2
14. mofa.gov.gh. (n.d.). Upper Denkyira East Municipal – Ministry of Food & Agriculture. Retrieved June 2, 2019, from http://mofa.gov.gh/site/?page_id=1468
15. Petersen W, Bertino L, Callies U, Zorita E. Process identification by principal component analysis of river water-quality data. Ecological Modelling. 2001;138(1-3):193-213. https://doi.org/10.1016/S0304-3800(00)00402-6
16. Rencher CA. Methods of Multivariate Analysis. In The Mathematical Gazette Second ed., 2002, 56. https://doi.org/10.2307/3613737.
17. Shweta T, Bhavtosh S, Prashant S, Rajendra D. Water quality assessment in terms of Water quality index. 2013. Retrieved May 27, 2019, from America Journal of water Resources website: https://www.google.com/search?source=hp&ei=IYnrXML6D4qQkwWohaCQDQ&q=Shweta%2C+T.%2C+Bhavtosh%2C+S.%2C+Prashant%2C+S.%2C+%26+Rajendra%2C+D.+%282013%29.+Water+quality+assessment+in+terms+of+Water+quality+index.+America+Journal+of+water+Resources.pdf&oq=Shwe
18. wikiversity.org. (n.d.). Exploratory factor analysis/Assumptions - Wikiversity. Retrieved May 31, 2019, from https://en.wikiversity.org/wiki/Exploratory_factor_analysis/Assumptions
19. Zhong L, Hope JG, Taylor TP, Gallo C, Pennell KD, Carriere PPE. Groundwater Quality. Water Environment Research. 2006;71(5):973-1053. https://doi.org/10.2175/106143099x133947