Application of Surface Water Quality Classification Models Using Principal Components Analysis and Cluster Analysis

Mohamed Ahmed Reda Hamed, Assistant Professor, Civil Engineering Department, Canadian International Colleague (CIC), El Sheikh Zayed, Giza, Egypt.
Email: moha_hamed@cic-cairo.com

ABSTRACT
Water quality monitoring has one of the highest priorities in surface water protection policy. Many variety approaches are being used to interpret and analyze the concealed variables that determine the variance of observed water quality of various source points. A considerable proportion of these approaches are mainly based on statistical methods, multivariate statistical techniques in particular. In the present study, the use of multivariate techniques is required to reduce the large variables number of Nile River water quality upstream Cairo Drinking Water Plants (CDWPs) and determination of relationships among them for easy and robust evaluation. By means of multivariate statistics of principal components analysis (PCA), Fuzzy C-Means (FCM) and K-means algorithm for clustering analysis, this study attempted to determine the major dominant factors responsible for the variations of Nile River water quality upstream Cairo Drinking Water Plants (CDWPs). Furthermore, cluster analysis classified 21 sampling stations into three clusters based on similarities of water quality features.

The result of PCA shows that 6 principal components contain the key variables and account for 75.82% of total variance of the study area surface water quality and the dominant water quality parameters were: Conductivity, Iron, Biological Oxygen Demand (BOD), Total Coliform (TC), Ammonia (NH₃), and pH.

However, the results from both of FCM clustering and K-means algorithm, based on the dominant parameters concentrations, determined 3 cluster groups and produced cluster centers (prototypes). Based on clustering classification, a noted water quality deteriorating as the cluster number increased from 1 to 3, thus the cluster grouping can be used to identify the physical, chemical and biological processes creating the variations in the water quality parameters.

This study revealed that multivariate analysis techniques, as the extracted water quality dominant parameters and clustered information can be used in reducing the number of sampling parameters on the Nile River in a cost effective and efficient way instead of using a large set of parameters without missing much information. These techniques can be helpful for decision makers to obtain a global view on the water quality in any surface water or other water bodies when analyzing large data sets especially without a priori knowledge about relationships between them.

Keywords:
Keyword: Surface water, Principal component analysis, Cluster analysis.
1. Introduction

The Nile constitutes the essential source of life in Egypt; it provides people with their fresh water needs. It is an essential factor of production and vital for agriculture, transport, tourism and henceforth the socio-economic development of the country. However, the Nile has become, to a great extent, adversely affected by human activities. On the other hand, industrial waste discharge, leakage of sewage by urban agglomeration and agricultural runoff contributes to the Nile contamination (Abd El-Daiem, S., 2011).

Surface water quality deterioration at the intakes of Cairo water treatment plants along River Nile due to increasing level of some pollutants concentration above the guidelines paid the attention of public concern and may cause health hazards. Thus, the need for better management of Cairo treatment plants water sources quality is becoming essential.

The water quality can be characterized by many parameters that represent a water composition in specific localities and time. Real raw data are mostly huge, it means that they are not normally distributed, often co-linear or auto correlated, including outliers or errors etc. Due to this reason, multivariate methods such as principal component analysis, cluster analysis, the factor analysis, and the discriminant analysis, are used, (Zeng and Rasmussen, 2005).

Principal Components Analysis (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 (factors or pc’s). These factors are ordered so that the first few retain most of the variation present in all of the original variables.

In cluster analysis, the objects are grouped on the basis of similarities within a class and dissimilarities among different classes (Panda et al. 2006). The similarities and dissimilarities are obtained on the basis of distance measures which are Euclidean and Manhattan, (Kaufman and Rousseeuw 1990).

Fuzzy clustering generalizes partition clustering methods by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual is a member of only one cluster. Fuzzy C-means (FCM) can be achieved through more careful and informed initialization based on data content. By carefully selecting the cluster centers in a way which disperses the initial cluster centers through the data space, the resulting FCM approach samples starting cluster centers during the initialization phase. The cluster centers are well spread in the input space, resulting in both faster convergence times and higher quality solutions.

K-means can be used for cauterizing monitoring stations with similar water quality characteristics. K-means cluster analysis is a divisive clustering method with k number of groups set a priori to analysis (Akume and Weber, 2002). Once the number of clusters is set as an input and cluster centroids are initialized, observations are added iteratively to the most similar cluster, whose centroid is then recalculated until all of the observations are grouped (Davis, 2002).

2. Study area

The Nile River enters Egypt at its southern boundary with Sudan and runs through 1000 km long narrow valley, then divided at a distance of 25 km north of Cairo into two branches (Rosetta and Damietta) forming a delta which ends at the Mediterranean
Cairo, located on the Nile River south of the Mediterranean Sea, just upstream of the point where the river widens into the Delta. Cairo has an area of 353 km2 with an average reach length along the river about 50 km (from Km 900 to km 950 referenced to Aswan High Dam). Figure (1) illustrates the study area layout which covers Cairo governorate along the River Nile, bounded by El Saff town (Giza Governorate) at Km 877.00 from the South and El Kanater town (Qalubia Governorate) at Km 953.00 from the North. The study scope will focus on the upstream of drinking water plants located in Cairo governorate along Nile River (Tibeen, Kafr Elw, North Helwan, Maadi, Fostat, El Roda and Rod Farg).

Figure (1) Study Area Layout

2.1 Cairo Drinking Water Plants (CDWPs)
Cairo water company (CWC), a subsidiary of the Holding Company of Water and Wastewater, produces potable water with an amount reaches to 6 million m3/day used by inhabitants of Greater Cairo, (CWC, Annual Report, 2018). This is done through 13 Cairo drinking water plants (Tibeen, Kafr Elw, North Helwan, Maadi, Fostat, El Roda,
Rod El Farg, Amerea, Mostrod, El Marg, El Obour, El Asher, Shubra el Khiema) distributed in Greater Cairo. Table (1) shows the annual average raw water, treated water and sludge & washing water for Greater Cairo drinking water plants, (CWC, Central laboratory annual technical report, 2018)

Table (1) CDWPs surface water source, annual average raw water and treated water

| Drinking water plant | Plant intake geographic location | Surface water source | Raw water production (m3/day) | Treated water production (m3/day) |
|----------------------|---------------------------------|----------------------|------------------------------|----------------------------------|
| Tibeen               | Cairo                           | River Nile           | 178608                       | 155649                           |
| Kafr Elw             |                                 |                      | 78238                        | 70728                            |
| North Helwan         |                                 |                      | 321003                       | 283539                           |
| Maadi                |                                 |                      | 209179                       | 161772                           |
| Fostat               |                                 |                      | 1114381                      | 1046974                          |
| El Roda              |                                 |                      | 323216                       | 164625                           |
| Rod El Farg          |                                 |                      | 819695                       | 720908                           |
| Amerea               |                                 | Ismailia Canal       | 404226                       | 389853                           |
| Mostrod              |                                 |                      | 1281328                      | 1155899                          |
| El Marg              |                                 |                      | 650000                       | 526232                           |
| El Obour             | New Cities                      |                      | 860000                       | 790000                           |
| El Asher             |                                 |                      | 600000                       | 500000                           |
| Shubra el Khiema     | Qalubia                         | Sharkawia Canal      | 379146                       | 358091                           |

From table (1) and according to the study scope which focus on seven CDWPs located on Cairo governorate along Nile River only (Tibeen, Kafr Elw, North Helwan, Maadi, Fostat, El Roda and Rod Farg).

3. Materials and Methods
3.1 Data Requirements
Surface Water samples were collected from various sampling locations of rivers, canal, drains and industrial pollution sources of the study area. The analyses of water samples were carried on twenty water quality parameters according to the standard methods for the examination of water and wastewater (APHA, 2012) for twelve consequence months during two years (2017, 2018) to show the effect of the spatial and temporal variation.

These water quality parameters included: pH, Turbidity, Electric Conductivity(EC), Total hardness, Total dissolved solids (TDS), Total alkalinity, Sulfates, Chlorides, Ammonia (NH₃), Nitrates (NO₃), Nitrites(NO₂), Phosphate, Iron, Manganese, Calcium, Magnesium, Aluminum, Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen(DO), Total Coliform(TC).

3.2 Methods
The methods consisted of three main components as follows:-
1) Perform the principal component analysis
2) Develop the dominant water quality parameters
3) Develop the optimum number of clusters by using Fuzzy C-Means (FCM)
4) Apply K-means algorithm technique to produce the generalized characteristics of clusters using the dominant parameters normalized data.
3.2.1 Principal component analysis

Principal component analysis (PCA) is mainly applied for the removal of data noise by the reduction of their dimensionality (Jolliffe, 2002). PCA searches new abstract orthogonal principal components (eigenvectors) which explain most of the data variation in a new coordinate system. Each principal component (PC) is a linear combination of the original variables and describes a different source of variation.

\[ PC_i = w_1x_1 + w_2x_2 + \ldots + w_nx_n \]  

(1)

Where \( x_i \) and \( w_i \) are the original variable and the component weight, respectively. The principal component weights are used as measures of the correlation between the variables and the principal components. The largest or first PC is oriented in the direction of largest variation of the original variables and passes through the center of the data. The second largest PC lies in the direction of the next largest variation, passes through the center of the data and is orthogonal to the first PC. The third largest PC is directed towards the next largest variance, goes through the data center and is orthogonal to the first and second PCs, and so forth. Classical PCA is based on the decomposition of a covariance/correlation matrix by the eigenvalue decomposition or by the singular value decomposition of real data matrices. The eigenvalues or singular values indicate variations among the observed variables, (Yu et al. 2003).

Before the computation, the testing data were standardized in order to avoid misclassifications arising from different orders of magnitude of tested variables. Therefore the original data were meaning (average) centered and scaled by the standard deviations.

Procedural steps of the PCA, (Suhr 2005) are:-

- Number of components equal to number of variables is generated
- The number of components to retain is determined
- Components are rotated (rotations is a linear transformation of the solution to make interpretation easier)
- Rotated solution is interpreted

3.2.2 Dominant water quality parameters

In this study, to determine the main dominant water quality parameter, Varimax rotation used as an effective orthogonal rotation method that minimized the number of variables that have high loading on each factor. The Varimax coefficient having correlation greater than 0.75 are considered as strong and indicate high proportion of its variance explained by the factor, between 0.50 and 0.75 is considered as moderate loading while 0.30-0.50 as weak significant factor loading, indicating much of that attribute’s variance remains unexplained and it is less important (Reghunath, et al 2002).

3.2.3 Fuzzy C-Means clustering (FCM) analysis

FCM applied for clustering the raw data into several categories using the selected operators without respect to any predetermined criteria in relation to each category. Most of the rules designed for FCM are based on the proper search for centroids or representative objects around which all observations will be clustered on a minimum basis (Selim, 1984; Trauwaert et al., 1991).

FCM seeks to minimize the following objective function, C, made up of cluster memberships and distances (Kaufman, 1990).

\[ C = \sum_{k=1}^{K} \sum_{j=1}^{N} m_{jk}^d d_{ij} \]

\[ + \sum_{j=1}^{N} m_{j}^d \]

Electronic copy available at: https://ssrn.com/abstract=3364401
Where \( m_{ik} \) represents the unknown membership of the object \( i \) in cluster \( k \) and \( d_{ij} \) is the dissimilarity between objects \( i \) and \( j \). The memberships are subject to constraints that they all must be non-negative and that the memberships for a single individual must sum to one.

One of the most difficult tasks in cluster analysis is to choose the appropriate number of clusters. In fuzzy clustering, the following coefficients are used:

I- The amount of ‘fuzziness’ in a solution may be measured by Dunn’s partition coefficient which measures how close the fuzzy solution is to the corresponding solution. This hard solution is formed by classifying each object into the cluster which has the largest membership. The formula for Dunn’s partition coefficient is:

\[
F(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_{ik}^2
\]

This coefficient ranges from \( 1/K \) to 1. Its value is \( 1/K \) when all memberships are equal to \( 1/K \). The value of one results when, for each object, the value of one membership is unity and the rest are zero.

II- Dunn’s partition coefficient may be normalized so that it varies from 0 (completely fuzzy) to 1 (hard cluster). The normalized version is

\[
F_{c}(U) = \frac{F(U) - (1 / K)}{1 - (1 / K)}
\]

III- Another partition coefficient, given in Kaufman (1990) is:

\[
D(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} (m_{ik} - m_{ik})^2
\]

IV- The normalized version of this equation is:

\[
D_{c}(U) = \frac{D(U)}{1 - (1 / K)}
\]

\( F_c (U) \) and \( D_c (U) \) together give a good indication of an optimum number of clusters. We should choose \( K \) so that maximize the value of \( F_c (U) \) and minimizes \( D_c (U) \) (Kaufman, 1990).

### 3.2.4 K-means algorithm

K-means is a simple and efficient algorithm. It divides \( n \) observations into given \( K \) clusters and each observation belongs to cluster with nearest mean. It uses the sum of square error criteria. The cluster pattern is assigned when sum of square error is minimum. The sum of square error equation (SSE) for K-means is given by (Kanungo et al. 2002):

\[
SSE = \sum_{C_i} \sum_{x \in C_i} ||x - m_i||^2
\]

Where \( m_i \) is the mean of the \( i^{th} \) cluster and \( x \in C_i \) is a pattern assigned to that cluster. The K-means clustering has advantage over other methods as it can be used to assign new cases to the existing clusters.
4. Results and Discussion

4.1 Descriptive Statistics

Basic statistics were carried out in order to give initial information about the water quality data. Table (2) shows the details of descriptive statistics for the water quality variables measured in two years.

Table (2) Mean and standard deviation of surface water quality parameters

| Variables | Mean | Standard Deviation | Minimum | Maximum |
|-----------|------|--------------------|---------|---------|
| PH        | 8.291| 0.041              | 8.200   | 8.450   |
| Turbidity | 7.751| 2.294              | 2.810   | 13.960  |
| EC        | 420.622| 59.143           | 338.000 | 584.000 |
| Total Hardness | 128.623| 12.188          | 116.000 | 164.000 |
| TDS       | 278.514| 40.173           | 176.500 | 385.440 |
| Total Alkalinity | 151.420| 7.131        | 138.000 | 170.000 |
| Sulphates | 24.679| 8.608             | 11.770  | 48.280  |
| Chlorides | 25.226| 7.461             | 11.520  | 51.000  |
| Ammonia(NH3) | 0.150| 0.090         | 0.010   | 0.470   |
| Nitrates(NO3) | 0.336| 0.311          | 0.010   | 1.220   |
| Nitrites(NO2) | 0.035| 0.051          | 0.010   | 0.490   |
| Iron      | 0.208| 0.185             | 0.003   | 0.690   |
| Manganese | 0.016| 0.017             | 0.002   | 0.064   |
| Calcium   | 31.470| 2.696            | 22.800  | 38.400  |
| Magnesium | 11.934| 1.778            | 0.030   | 17.280  |
| Aluminium | 0.052| 0.037             | 0.001   | 0.140   |
| BOD       | 3.851| 0.202             | 3.370   | 4.120   |
| COD       | 17.013| 0.177            | 16.580  | 17.470  |
| DO        | 7.473| 0.153             | 2.000   | 7.990   |
| TC        | 20106.330| 3327.820      | 14000   | 24000   |

4.2 Principal Component Analysis

The calculated principal components loadings, eigenvalues, total variance and cumulative variance are shown in Table(3) while the scree plot of the eigenvalues of observed components is depicted in Figure (2).
Table (3) Principal component analysis after Varimax Rotation

| Variables     | PC1   | PC2   | PC3   | PC4   | PC5   | PC6   |
|---------------|-------|-------|-------|-------|-------|-------|
| PH            | 0.241 | 0.038 | -0.085| -0.385| -0.054| 0.701 |
| Turbidity     | -0.124| 0.455 | -0.065| 0.135 | 0.600 | 0.046 |
| EC            | 0.902 | 0.134 | -0.091| 0.165 | 0.105 | 0.050 |
| Total Hardness| 0.887 | 0.149 | 0.046 | -0.124| 0.121 | 0.059 |
| TDS           | 0.889 | 0.196 | -0.082| 0.154 | 0.131 | 0.059 |
| Total Alkalinity| 0.650 | -0.529| -0.132| 0.290 | 0.078 | 0.169 |
| Sulfates      | 0.833 | -0.115| -0.278| 0.071 | 0.036 | -0.016|
| Chlorides     | 0.818 | -0.454| -0.106| 0.147 | 0.038 | -0.001|
| Ammonia (NH3) | -0.137| -0.001| 0.043 | -0.130| 0.853 | -0.043|
| Nitrates (NO3)| -0.674| -0.122| -0.083| 0.170 | -0.430| -0.205|
| Nitrites (NO2)| -0.336| -0.183| -0.134| 0.362 | 0.028 | 0.256 |
| Iron          | -0.038| 0.879 | -0.075| -0.134| 0.158 | 0.116 |
| Manganese     | 0.088 | 0.819 | 0.007 | 0.090 | 0.038 | -0.018|
| Calcium       | 0.726 | 0.536 | 0.003 | -0.151| 0.088 | 0.075 |
| Magnesium     | 0.811 | -0.240| 0.087 | -0.079| 0.075 | 0.005 |
| Aluminum      | -0.332| 0.133 | 0.042 | 0.353 | 0.286 | 0.589 |
| BOD           | 0.099 | 0.045 | 0.938 | 0.065 | 0.009 | 0.065 |
| COD           | 0.096 | -0.099| 0.926 | 0.086 | 0.017 | -0.144|
| DO            | 0.353 | -0.002| 0.025 | -0.127| 0.062 | 0.666 |
| TC            | 0.066 | 0.003 | -0.167| 0.760 | 0.100 | 0.021 |
| % Variability | 31.480| 13.290| 9.730 | 6.550 | 7.360 | 7.400 |
| Cumulative percentage | 31.480 | 44.780 | 54.510 | 61.060 | 68.420 | 75.820 |

Figure (2) Scree plot of the eigenvalues of observed components

The results of principal components analysis illustrated in Table (3) and Figure (2) of Cattel scree plot (Cattel, 1966) show that of the 20 components, only 6 had extracted eigenvalues over 1. This is based on Chatfield and Collin (1980) assumption which stated that components with an eigenvalue of less than 1 should be eliminated. The extracted 6 components were subsequently rotated according to varimax rotation in order to make interpretation easier and fundamental significance of extracted components to the water quality status of the selected study period. The result of rotation revealed further, the percentages of the total variances of the 6 extracted components when added account for 75.82% (that is their cumulative variance) of the...
total variance of the observed variables. This indicates that the variance of the observed variables had been accounted for by these 6 extracted components.

As it is obvious, the first principal component (PC1), accounts for 31.48% from total variation, can be called as an indication of salt component because it is mainly saturated with conductivity, hardness (including calcium). PC1 accounts show a strong loading on EC (0.902), TDS (0.889), Total Hardness (0.887), Sulfates (0.883), Chlorides (0.881), Magnesium (0.811), while moderate loading on Calcium (0.726), Nitrates (0.674) and Total Alkalinity (0.65). Electric Conductivity(EC) measurements indicate the presence of dissolved salts and electrolytic contaminants, but it gives no information about specific ion compositions, (Adekunle et al., 2007). There was a strong positive correlation between TDS and EC values which revealed positively strong correlation to each other (r= +0.99), so the study results were in accordance with Toufeek and Korium (2009).

The second principal component (PC2), accounts for 13.29%, is associated with strong loading on Iron (0.879) and Manganese (0.819), while moderate loading on Calcium (0.536). The concentration of iron and manganese recorded higher attribute due to the intense of human activities and industrial effluents from iron and steel companies.

The third principal component (PC3) described 9.73% of the total variance had a strong positive loading on BOD (0.938) and COD (0.926). These factors loading explained the effects of organic pollution and reflect strong influence of anthropogenic activities in the area, probably from domestic waste and industrial waste. High BOD and COD levels in the study area are related to the existence of high bacterial load and organic matters as well as relative high temperatures which enhance the enumeration of bacteria. However all results of study area water samples were higher than the permissible limit guidelines (COD should not exceed 10mg/l) according to Egyptian National water quality standards, Law 48/1982 regarding the protection of the River Nile and waterways from pollution.

Out of the total variance, 6.55% is explained by the fourth principal component (PC4), is mainly carried by TC with a positive strong correlation (0.76) that is indicators for water contamination. The high counts of total coliform might be due to pollution by industrial activities discharging their wastes to the Nile water in Cairo (Saleh, 2009). All results of Nile water samples were higher than the permissible limit guidelines (TC should not exceed 5000 cfu/100 ml) according to Tebbutt (1998). The study results also agree with Sabae and Rabeh (2006).

Additionally, 7.36% of the total variance of water quality is exhibited by NH3 with a strong positive loading under the fifth principal component (PC5). NH3 is closely related to the organic matter contents of the sediment and this high amount of nutrients might also result from the application of manure in agricultural activities (Terceiro et al, 2008).

The six principal component (PC6), with 7.4% of the total variance, consists mainly of pH (0.701) and DO (0.701) with a moderate loading. This factor resulted due to the anaerobic conditions in the river from the strong loading of dissolved organic matter which leads in the formation of organic acids. pH value has an effect on the biological, chemical reactions, as well as it controls the metal ion solubility and thus it affects the natural aquatic life. The study results were in accordant with Toufeek and Korium (2009).
Based on the component loadings, the variables are grouped accordingly with their designated components as follows:

- Component 1: EC, TDS, Total Hardness, Sulfates, Chlorides, Magnesium, Calcium, Nitrates and Total Alkalinity.
- Component 2: Iron and Manganese.
- Component 3: BOD and COD.
- Component 4: TC.
- Component 5: NH$_3$.
- Component 6: pH and DO.

4.3 Dominant water quality parameters

The dominant parameters identified by the PCA are: EC, Iron, BOD, TC, NH3 and pH (see Table 3). The previous discussion indicated that most of measured water quality parameters such as EC, TDS, Total Hardness, different major ions and Total Alkalinity, loaded with positive values, and they have strong effects on PC1. EC has the maximum strong loading value in PC1. Thus, EC is considered as a dominant parameter.

The Iron is considered as the next dominant water quality parameter as it is loaded strong in PC2 with the highest positive values. Also, the concentrations iron in the Nile water cause the exceedance of the drinking water guidelines, particularly at the anthropogenic impact points, where, iron is regulated by secondary drinking water contaminant that may cause offensive taste, odor, color corrosion or staining problems, (EWQS, 2007).

The BOD is considered as the third dominant water quality parameter as it is loaded strong in PC3 with highest value (0.938). These two parameters (BOD and COD) may have a strong relationship with each other, particularly the discharge of industrial and agricultural effluents containing a large amount of organic matter, (Goher et al., 2015).

The TC is considered as the fourth dominant water quality parameter as it is loaded strong in PC4 with highest value (0.760).

The Ammonia is considered as the fifth dominant water quality parameter as it is loaded strong in PC4 with highest value (0.853). Ammonia may result from fertilizers that are present in soil and it is relatively easily oxidized to nitrite and finally to nitrate (Karavoltsos et al., 2008) and it possesses a serious threat to public health.

The pH is considered as the third dominant water quality parameter as it is loaded strong in PC3 with highest value (0.701). pH value has an effect on the biological, chemical reactions, as well as it controls the metal ion solubility and thus it affects the natural aquatic life. Moreover pH could control the pathogenic microorganism growth (Zamaxaka et al., 2004).

4.4 Cluster analysis

4.4.1 Optimum number of clusters

FCM applied to determine the optimum number of clusters (k) that maximize the value of Fc (U) and minimize Dc (U) (Kaufman, 1990). Table (4) illustrates the values of Fc (U) and Dc (U) with the corresponding number of clusters.
Table (4) Results of optimum number determination

| Number of clusters(K) | F(U)    | Dc (U)  |
|-----------------------|---------|---------|
| 2                     | 0.3333  | 0.9998  |
| 3                     | 0.2500  | 1.0122  |
| 4                     | 0.2000  | 1.0131  |
| 5                     | 0.1667  | 1.0134  |
| 6                     | 0.1429  | 1.0136  |
| 7                     | 0.1250  | 1.0137  |
| 8                     | 0.3333  | 0.9998  |

FCM results illustrated in the table (4), it noticed that the optimum number of clusters for the study area is three clusters which satisfies the above conditions.

4.4.2 Clusters characteristics

According to the optimum number of clusters which determined by using FCM in the previous step, K-means algorithm applied to produce the generalized clusters characteristics using the dominant parameters. After finding medians of clusters, the clusters are developed by assigning each object of dataset to the nearest medians of the clusters. The dissimilarities from each of the objects in the dataset from these centers of the clusters are determined using Euclidean distance. Cluster Centers are selected on the basis of the minimum distance. Silhouette is used for interpretation and validation of clusters, (Kaufman and Rousseeuw 1990). Table (5) and Figure (3) illustrate generalized characteristics mean values and the six dominant parameters mean values of the three clusters respectively.

Table (5) Clusters centers mean values

| Dominant Parameter | Cluster 1 | Cluster 2 | Cluster 3 |
|--------------------|-----------|-----------|-----------|
| PH                 | 8.303     | 8.340     | 8.4010    |
| Turbidity          | 8.052     | 8.5382    | 10.537    |
| EC                 | 414.611   | 502.888   | 539.111   |
| Total Hardness     | 148.444   | 153.200   | 164.666   |
| TDS                | 334.976   | 363.307   | 398.026   |
| Total Alkalinity   | 143.222   | 150.622   | 163.777   |
| Sulphates          | 23.484    | 25.683    | 27.358    |
| Chlorides          | 19.777    | 20.300    | 24.777    |
| Ammonia            | 0.225     | 0.120     | 0.228     |
| Nitrates           | 0.497     | 0.531     | 0.684     |
| Nitrates           | 0.118     | 0.126     | 0.217     |
| Iron               | 0.120     | 0.155     | 0.617     |
| Manganese          | 0.006     | 0.014     | 0.039     |
| Calcium            | 34.022    | 34.302    | 34.666    |
| Magnesium          | 10.854    | 11.141    | 11.300    |
| Aluminium          | 0.069     | 0.043     | 0.068     |
| BOD                | 3.792     | 3.852     | 3.944     |
| COD                | 16.975    | 17.006    | 17.950    |
| DO                 | 7.219     | 7.402     | 7.563     |
| TC                 | 18266.670 | 19922.220 | 21666.670 |

Electronic copy available at: https://ssrn.com/abstract=3364401
It is obvious from table (5) and figure (3) for the K-means algorithm results, as the cluster number changed from 1 to 3, the value of the six dominant parameters and the water quality deterioration increased.

4.4.3 DWP\(s\) and monitoring stations clusters allocation

According to the K-means algorithm generalized clusters characteristics results, the allocation for CDWPs and monitoring stations clusters were developed. The output of the cluster characteristics analysis is dispensed in dendogram, Figure (4). Dendogram gives the picture of the clusters describing the spatial variation in the water quality and the grouped monitoring stations, CDWPs of each cluster.
Based on the results of cluster analysis, stations and CDWPs grouped under each cluster in Figure (4), it was concluded that:-

The first cluster, mainly located in the upstream of the study area with less polluted (LP) stations, included the stations from (878.00) to (868.00) and three DWPs (Tibeen, Kafr Elw and North Helwan). The changes in water quality in this cluster were mainly due to the agricultural drainage water mixed with partially treated or untreated domestic wastewater, industrial wastewater and wastewater from these three drinking water plants sludge disposal. The second cluster, comprised only the three DWPs (Maadi, Fostat and El Roda) with moderate pollution (MP), is mainly affected by the cumulative pollution from the previous cluster in addition to the wastewater from the three mentioned drinking water plants sludge disposal. The common feature of these sites was relatively high dominant parameters concentrations compared to the first cluster. The third cluster located in the downstream of the study area, included Rod El Farag DWP and the different stations from (932.00) to (950.00) along the river, had the highest pollution level (HP). These stations are distinguished from other stations concerning the level of pollution and have the most distance from other stations.

5. Conclusions
This study revealed the usefulness of multivariate statistical techniques for analysis and interpretation of complex data sets, and in water quality assessment, understanding temporal variations in water quality management. The PCA revealed that three principal components (PCs) were able to explain ~75.82% of the variability and the dominant water quality parameters were total EC, Iron, BOD, TC, NH3 and pH.

In this study, a methodology for clustering twenty one monitoring station including seven CDWPs along Nile River using dominant water quality parameters. FCM indicated three as the most optimum number of clusters. K-means clustering technique is used on the normalized data of the dominant parameters to obtain the generalized characteristics of three clusters. The water quality deteriorated as the cluster number increased from 1 to 3. Pattern-match using K-means clustering technique was applied to allocate clusters to all monitoring station including seven CDWPs along Nile River for two consecutive years (2017 and 2018). The Cluster analysis confirmed the existence of three types of clusters water quality (i.e., low-, medium-, and high-polluted). According to the study methodology, the decision makers can develop optimal strategy in which sampling stations can be reduced, identifying the specific pollutants in source waters for designing economical, targeted and effective management for drinking water plants facilities.

5. References
- Abd El-Daiem, S., 2011. Water Quality Management in Egypt. Journal of Water Resources Development, 27(1): 181-202.
- Adekunle, L., Adetunji, M., Gbadebo, A. (2007). Assesment of ground water quality in a typical rural settlement in south Nigeria, Int. J Environ. Res. Public Health 4 (4): 307-318.
- Akumne, D. and Weber, G.-W. 2002. Cluster Algorithms: Theory and Methods. Journal of Computational Technologies, 7(1):15-27.
- Cairo Drinking Water Company, Central laboratory annual technical report, 2018.
- Cattel, R., D.: The scree test for the number of factors. Multivariate Behav. Res. 1, 1966, 245-276.
- Chatfield C, Collin AJ (1980). Introduction to Multivariate Analysis. Chapman and Hall in Association with Methuen, Inc. 733 Third Avenue, New York NY.

- Clare, A. A.; Dianne F. J. and Stuart L. S. (2007). Effect of overlying water pH, dissolved oxygen, salinity and sediment disturbances on metal release and sequestration from metal contaminated marine sediments, Chemosphere 69 (9) p 1428–1437.

- Davis, J.C., 2002. Statistics and Data Analysis in Geology (Third Edition). John Wiley and Sons, Inc., New York, NY.

- Egyptian Governmental Law No. 48, 1982. The Implementer Regulations for law 48/1982 regarding the protection of the River Nile and water ways from pollution. Map. Periodical Bull., 3-4: 12-35.

- EWQS (Egyptian drinking water quality standards), 2007. Ministry of Health, Population Decision number 458.

- Goher, M.E., Hassan, A.M., Abdel-Moniem, I.A., Fahmy, A.H., El- Sayed, S.M., 2014. Evaluation of surface water quality and heavy metal indices of Ismailia Canal, Nile River, Egypt. J. Aquat. Res. 40, 225–233.

- Karavoltsos, S., Sakellar, A., Mihopoulos, N., Dassenakis, M., Scoullos, M.J. (2008). Evaluation of the quality of drinking water in regions of Greece. Desalination, 224: 317-329.

- Jolliffe, I., T. Principal component analysis (2nd edn.). Springer-Verlag, New York, 2002.

- Kaufman L., Rousseeuw P. J., Finding Groups in Data—An Introduction to Cluster Analysis, John Wiley & Sons Inc., New York 1990.

- Laurie, K., Bryan, F. and manly, J. Manly (2005). Multivariate Statistical Methods: A Primer .Chapman & Hall/Crc .

- McKenna, J.E. (2003). An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environmental Modelling & Software, 18 (3): 205-220.

- Panda U. C., Sundaray S. K., Rath P., Nayak B. B., Bhatta D., Application of Factor and Cluster Analysis for Characterization of River and Estuarine Water Systems – A Case Study: Mahanadi River (India), J. Hydrol. 2006, 331, 434-445.

- Rabeh, S.A. (2001). Ecological stufies on nitrogen cycle bacteria in Lake Manzalah, Egypt. Egypt. J. Aquat. Biol. & Fish., 5(3): 263 -282.

- Reghunath R, Murthy S. T. R & Raghavan B. R (2002). The utility of multivariate statistical techniques in hydrogeochemical studies. An example from Karnataka, India, Water Research 36. 2437-2442.

- Saleh, A.R. (2009). Bacteria and viruses in the Nile. Monographiae Biologicae, 89: 407-429.

- Selim, S. Z. (1984). Soft clustering of multi-dimensional data: a semi-fuzzy approach. Pattern Recognition 17(5), 559–568.

- Suhr D (2005) Principal component analysis vs. exploratory factor analysis. SUGI 30 Proceedings. Retrieved January 30 from: http://www2.sas.com/proceedings/sugi30/Leads30.pdf. Quality in Lake Nasser water. Global J. Environ. Res., 3 (3): 141- 148.

Electronic copy available at: https://ssrn.com/abstract=3364401
- Terceiro, P., Lobo-Ferreira, J. P., & Leitão, T. E. (2008). Análise da qualidade da água e questões de governança Albufeira Alqueva. Comunicação apresentada no 9º Congresso da Água – Água: Desafios de hoje, exigências de amanhã. Cascais, Portugal. http://www.aprh.pt/congressoagua2008/PDF/Lobo-Ferreira Alqueva.pdf. Accessed 20 January 2009 (in Portuguese).

- Tebbutt, T. (1998). Principles of Water Quality Control. 5th Ed., Hallam University.

- Toufeek, M.A. and Korium, M.A. (2009). Physico-chemical characteristics of water quality in Lake Nasser water. Global J. Environ. Res., 3 (3): 141-148.

- Trauwaert, E., Kaufman, L., Roussseeuw, P. (1991). Fuzzy clustering algorithms based on the maximum likelihood principle. Fuzzy Sets and Systems 42, 213–227.

- Toufeek, M.A. and Korium, M.A. (2009). Physico-chemical characteristics of water.

- WHO (1989) Health Guidelines for the Use of Wastewater in Agriculture and Aquaculture. Report of a WHO Scientific Group, Technical Report Series No. 778, WHO, Geneva.

- Yu S, Shang J, Zhao J, Guo H (2003) Factor analysis and dynamics of water quality of the Songhua River Northeast China. Water Air and Soil Pollut 144: 159–169

- Zeng, X., Rasmussen, T.C.: Multivariate statistical characterization of water quality in lake Lanier, Georgia, USA. J. Environ. Qual. 34 (6), 2005, 1980-1991.

- Zaxaka, M., Pironcheva, G. and Muyima NYO. (2004). Microbiological and physico-chemical assessment of the quality of domestic water sources in selected rural communities of the Eastern Cape Province, South Africa.