Factor Analysis as a Tool to Identify Water Quality Index Parameters along the Nile River, Egypt

Mohsen M. Yousry[1] and H.A.A El Gammal[2]

[1]Nile Research Institute (NRI), National Water Research Center, Cairo, Egypt
[2] National Water Research Center (NWRC), Cairo, Egypt
elgammalhussein@gmail.com

Abstract: Water quality is of great importance in managing water as it is directly affects humans health. The Nile River and its two branches are the main sources of water in Egypt. They play a major role in assimilation or transportation of the municipal and industrial wastewater discharge as well as the occasional or seasonal pollution sources. In this research, the factor analysis technique is applied to surface water quality data sets collected from the Nile River to determine the sources of pollution during two different hydrological periods (low and high flow conditions) as well as to identify water quality index parameters. Data were collected for two years from 2010 to 2012 for both flow conditions and were analyzed by descriptive statistics. The results showed that all water quality parameters have mean values within the allowable limits except the chemical oxygen demand (COD) comparing to the national standards. Factor analysis revealed that water quality of the River was strongly affected by agricultural uses. On the other hand, the main pollution source changed from agricultural uses to agricultural uses mixed with organic contamination discharging from domestic wastewater in high-flow periods. The bacterial contamination represents second and the third factor during August and February; respectively. In addition, the organic contamination represents the last factor during the low flow period. This technique is very useful to decision makers in identifying priorities to improve water quality that has deteriorated due to the wastewater discharge.
[Mohsen M. Yousry and H.A.A El Gammal. **Factor Analysis as a Tool to Identify Water Quality Index Parameters along the Nile River, Egypt.** *J Am Sci* 2015;11(2):36-44]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 5

Key words: Nile River**,** factor analysis**,** water quality, agricultural and bacterial contaminations

1.Introduction

The surface water quality within a region is governed by natural processes (precipitation rate, weathering, and soil erosion) and anthropogenic effects (urban, industrial and agricultural activities and the human exploitation of water resources) **Nouri *et al.* (2008**). The possible variances in water quality may be due to anthropogenic activities, various biochemical (natural variances during season), or chemical processes.

Water quality monitoring has one of the highest priorities in environmental protection policy. Its main objective is to control and minimize the incidence of pollutant-oriented problems, and to provide water of appropriate quality to serve various purposes such as potable and irrigation water. The water quality is identified in terms of its physical, chemical and biological parameters (**Sargaonkar and Desh-pande, 2003**). The particular problem in the case of water quality monitoring is the complexity associated with analyzing a large number of measured variables.

The Nile River in Egypt flows downstream from High Aswan Dam (HAD) northward for about 950 km to a point (Delta Barrage) where it is divided into two branches. Monitoring program results in a huge and complex data matrix consist of a large number of physical and chemical parameters. The gathered data set is subjected to the multivariate statistical methods, which have been used successfully in hydrochemistry for many years. Surface water, groundwater, quality assessment, and environmental research employing multi-component techniques are well described in the literature (**Praus, 2005**). Multivariate statistical approaches allow deriving hidden information from the data set about the possible influences of the environment on water quality. Factor analysis attempts to explain the correlations between the observations in terms of the underlying factors, which are not directly observable. According to **Gupta *et al.* (2005)**, there are three stages in factor analysis. For all variables, a correlation matrix is generated factors which extracted from the correlation matrix based on the correlation coefficients of the variables. To maximize the relationship between some of these factors and variables, the factors are rotated.

The first step is the determination of the parameter correlation matrix. It is used to account for the degree of mutually shared variability between individual pairs of water quality variables. Then, Eigen-value and factor loadings for the correlation matrix have been determined. Eigenvalues correspond to an Eigen-factor which identifies groups of variables that are highly correlated among them. Lower Eigen-value may contribute little to the explanatory ability of the data. Factor rotation is used to facilitate

interpretation by providing a simpler factor structure (**Zeng and Rasmussen, 2005**).

There are more than 34 measured water quality parameters in lab and field two times in a year (NRI, 2012). Significant reduction of the measured parameters based on descriptive statistics adds an economic value (cost reduction) to the water quality monitoring program. A water quality index (WQI) is a tool to summarize large amounts of water quality data into simple terms (e.g. good or fair) for reporting to the public and decision makers in a consistent manner. It evaluates and ranks the quality of water bodies for various beneficial uses of water, such as habitat for aquatic life, agricultural irrigation and livestock water, recreation and aesthetics, and drinking water supplies. The detailed formulation of the WQI is described in the Canadian Water Quality Index 1.0 – Technical Report (**CCME, 2001**).

The main objective of the research is to define the main contributing pollution source to River Nile in low and high season flow. In addition, identify the most important water quality parameters that can be used in water quality indices.

2.Methodology

According to national water quality program, Nile Research Institute (NRI), National Water Research Center (NWRC), is responsible for monitoring the water quality along Aswan High Dam Lake, Nile reaches and branches. Also, it is responsible for monitoring agricultural drains, and main irrigation canals in Upper Egypt. The monitoring program is carried two times yearly during low and high flow conditions (in February and August, respectively). Water samples are collected for two years from 2010 to 2012 for both flow conditions from 18 sites along the Nile reach from Aswan to Delta barrage, as shown in figure (1).

Descriptive statistics are used to present quantitative descriptions in a manageable form. They provide simple summaries about samples and the measurements of physical, chemical and biological parameters that analyzed according to A**PHA (1998)**.

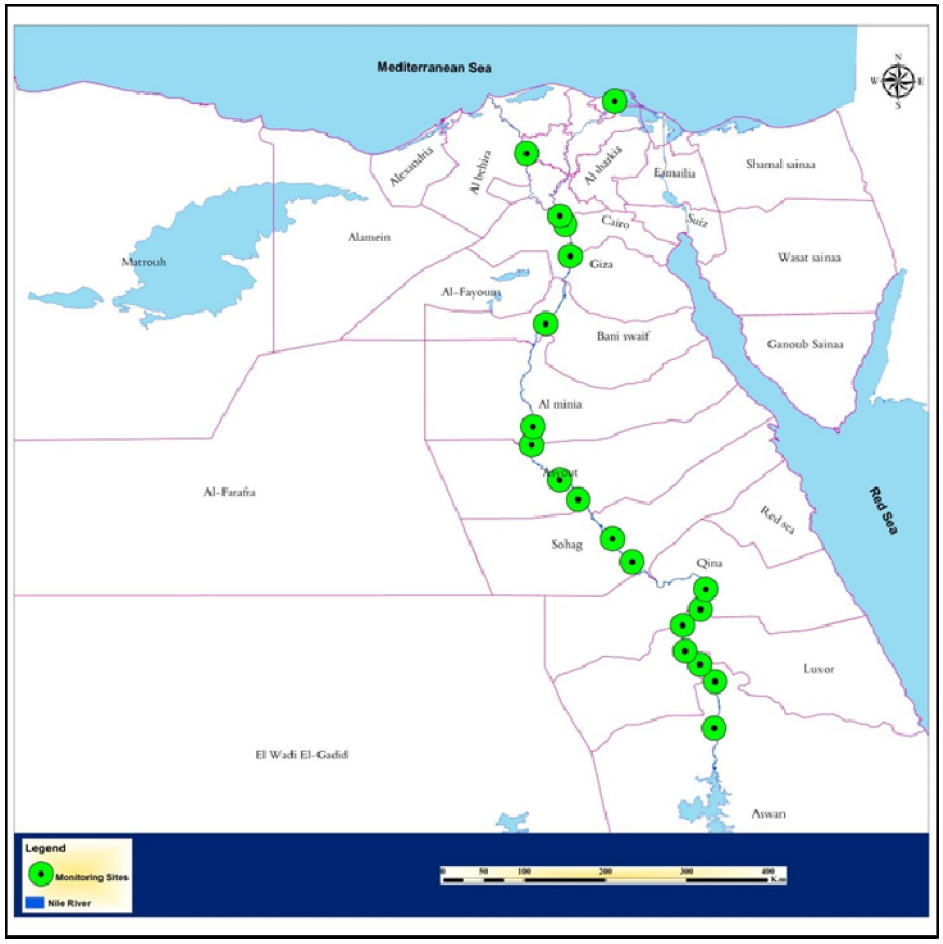


Figure (1): Nile River monitoring sites

Normality tests are used to determine whether a random variable is normally distributed or not. Many data analysis methods (e.g. t-test, ANOVA, regression, or multivariate techniques) depend on the assumption that data are normally distributed.

Singh *et al.* (2004 and 2005) provided information to describe the collected data. They clarified that Principal Component Analysis/Factor Analysis (PCA/FA) is a very powerful technique which provides information on the meaningful parameters to describe the whole data set rendering data reduction with minimum loss of information. In addition, **Kowalkowski *et al.* and Boyacioglu (2006)** explained that the PCA/FA is a quantification of the significance of variables that explain the observed grouping and patterns of the inherit properties of the individual objects and allows an explaining of related parameters by only one factor which responsible for variation in river water quality and eventually leads to sources identification of river water pollution.

In this research, PCA/FA was applied to extract the most significance PC's and to reduce the contribution of less significant variables to simplify even more of the data structure coming from PCA/FA. The obtained PC's were further subjected to varimax rotation according to well established rules to maximize differences between variables and facilitate smooth interpretation of the data (**Shrestha and Kazama, 2007**). The rotating axis is defined by PCA/FA generates Varimax Factor (VF) through factor analysis which can further reduce the contribution of variable with minor significance.

One of the most important steps of factor analysis is to determine number of factors that need to be extracted for accurate data analyses. In this regard, the rotation of the factor axis is performed to enhance a simple structure such as factors indicated by high loadings for some variables and low loadings for the others. In order to determine the number of factors to be used, variances and co-variances of the variables are computed. Then, eigenvalues and eigenvectors could be evaluated for the covariance matrix and the data is transformed into factors.

All mathematical and statistical calculations were implemented using Minitab and SPSS software packages.

3.Results and Discussions

Descriptive statistics and normality test

In this research, the descriptive statistics are conducted on the water quality data and the results indicated that the normality is not achieved but when the data are transformed into natural logarithm (ln) the normality was achieved, as shown in sample figures (2, 3 and 4). It's worth mentioning that, the normal distribution is indicated if the mean and the median of the data set are nearly equal.

By comparing the mean value of each variable with the national standards of Law 48/year 1982 and its ministerial decree 92 year 2013, it was found that all water quality variables have mean values within the permissible limits except chemical oxygen demand (COD). Its recorded values are slightly above the allowable limits recommended by national standards (10 mg/l). The high values of COD along the River Nile indicate that it receives a considerable load of organic matter (Table 1, 2). It is worth to mention that all values are in (mg/l) except turbidity (NTU), fecal coliform (CFU) and pH (unit).

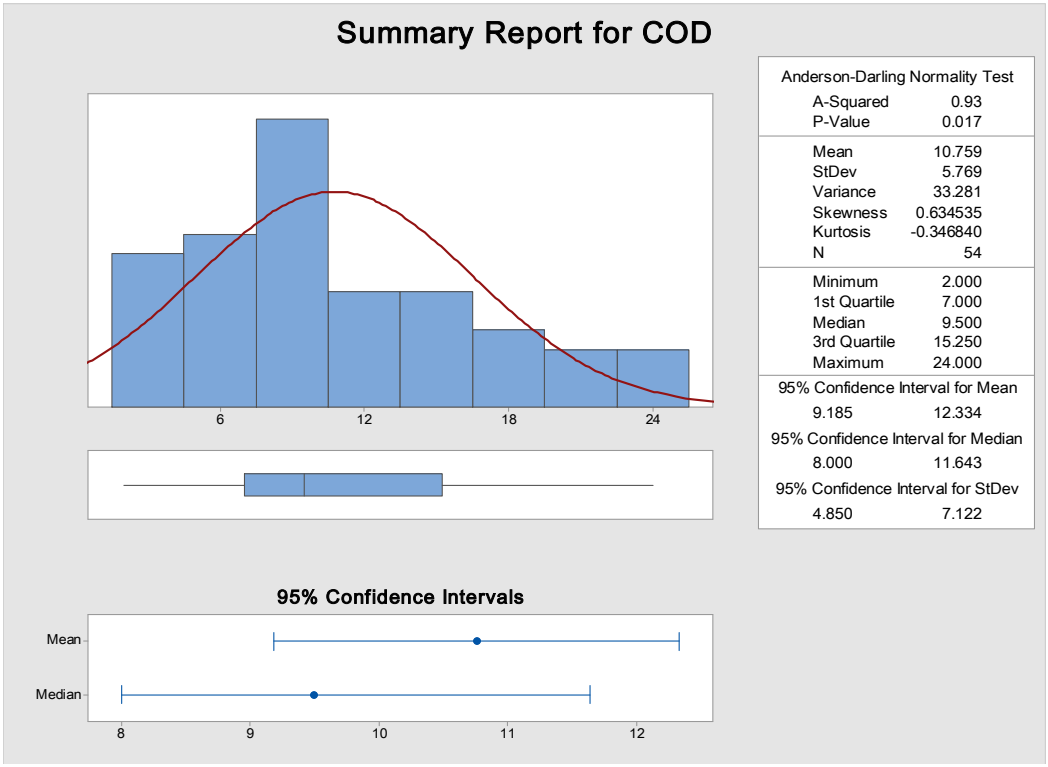
Table (1): Descriptive statistics of parameters during February

Statistic	DO	pH	Turbidity	Fecal Coliform	Nitrate	Ortho-Phosphate	Total-Phosphorus	BOD	COD	TSS	TDS
Mean	8.16	8.30	10.56	526.24	0.22	0.031	0.057	3.57	10.76	12.00	192.09
Median	8.23	8.26	9.83	76.50	0.21	0.030	0.060	3.00	9.50	11.00	192.00
Variance	0.26	0.02	26.88	3552936.98	0.01	0.000	0.001	7.67	33.28	26.23	752.73
Std. Deviation	0.51	0.16	5.18	1884.92	0.09	0.015	0.025	2.77	5.77	5.12	27.44
Minimum	7.10	8.05	2.62	1.00	0.00	0.006	0.011	1.00	2.00	3.00	149.00
Maximum	9.20	8.65	23.90	13000.00	0.49	0.083	0.120	17.00	24.00	25.00	239.00
Range	2.10	0.60	21.28	12999.00	0.49	0.077	0.109	16.00	22.00	22.00	90.00
Law	Not less than 6 mg/l	6.5-8.5			2 mg/l as N		2 mg/l	6 mg/l	10 mg/l		500 mg/l

Table (2): Descriptive statistics of parameters during August

Statistic	DO	pH	Turbidity	Fecal Coliform	Nitrate	Ortho-Phosphate	Total-Phosphorus	BOD	COD	TSS	TDS
Mean	7.55	8.19	10.46	114.24	.37	.032	.054	3.64	10.37	12.11	193.56
Median	7.88	8.17	9.56	40.00	.21	.030	.050	3.00	9.50	10.00	187.00
Variance	0.66	0.09	30.73	40063.66	.28	.000	.001	4.17	23.33	36.33	978.10
Std. Deviation	0.81	0.30	5.54	200.16	.53	.012	.024	2.04	4.83	6.03	31.27
Minimum	5.40	7.51	2.64	0.00	.00	.014	.015	1.17	3.00	3.00	151.00
Maximum	8.74	8.80	24.30	1200.00	3.21	.070	.110	11.00	24.00	26.00	274.00
Range	3.34	1.29	21.66	1200.00	3.21	.056	.095	9.83	21.00	23.00	123.00
Law	Not less than 6 mg/l	6.5-8.5			2 mg/l as N		2 mg/l	6 mg/l	10 mg/l		500 mg/l

Factor analysis



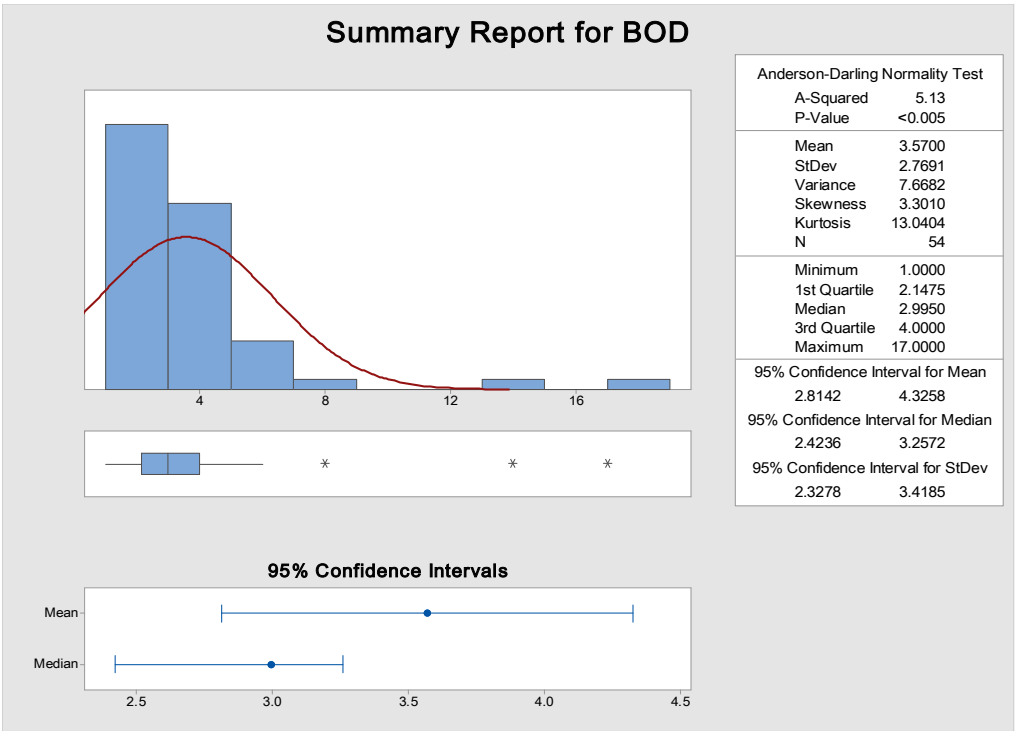
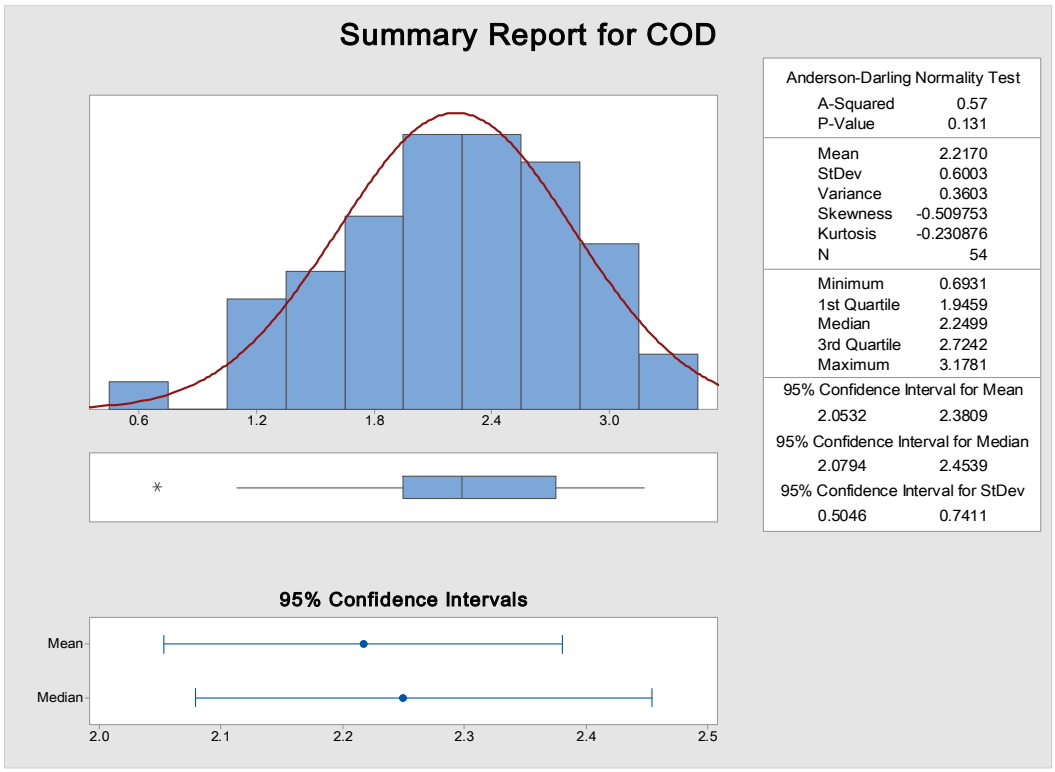


Figure (2): Histogram, boxplot and confidence intervals of some variables (raw values) along the Nile during February (2010-2012)



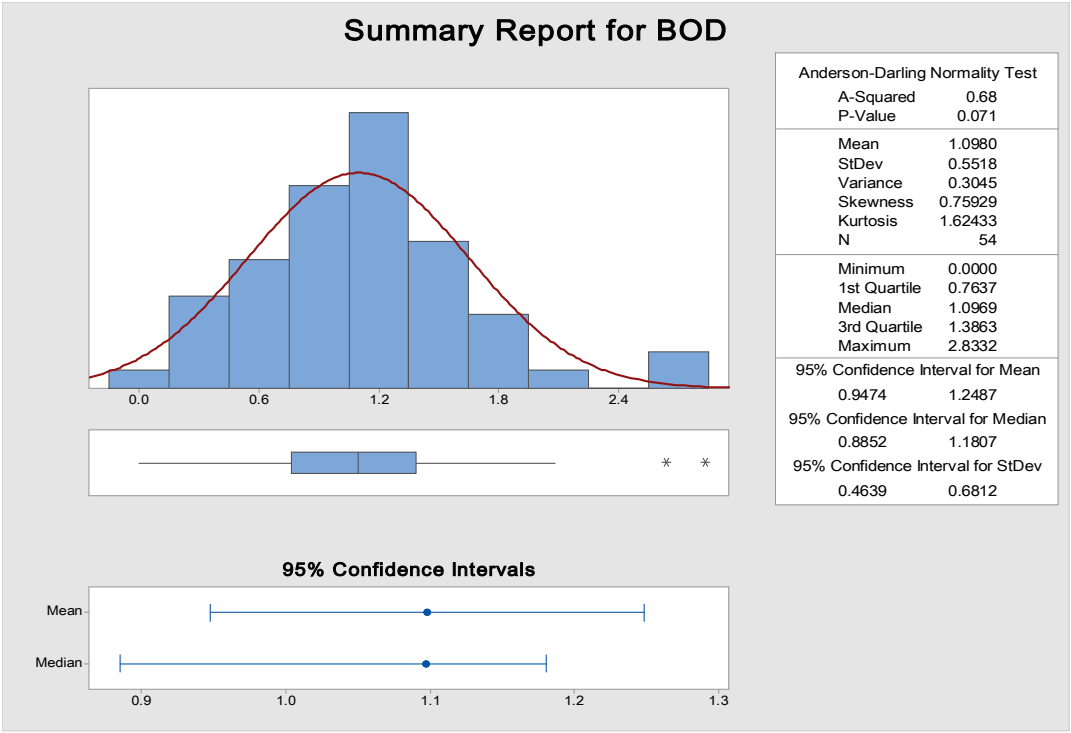


Figure (3): Histogram, boxplot and confidence intervals of some variables (ln values) along the Nile during February (2010-2012)

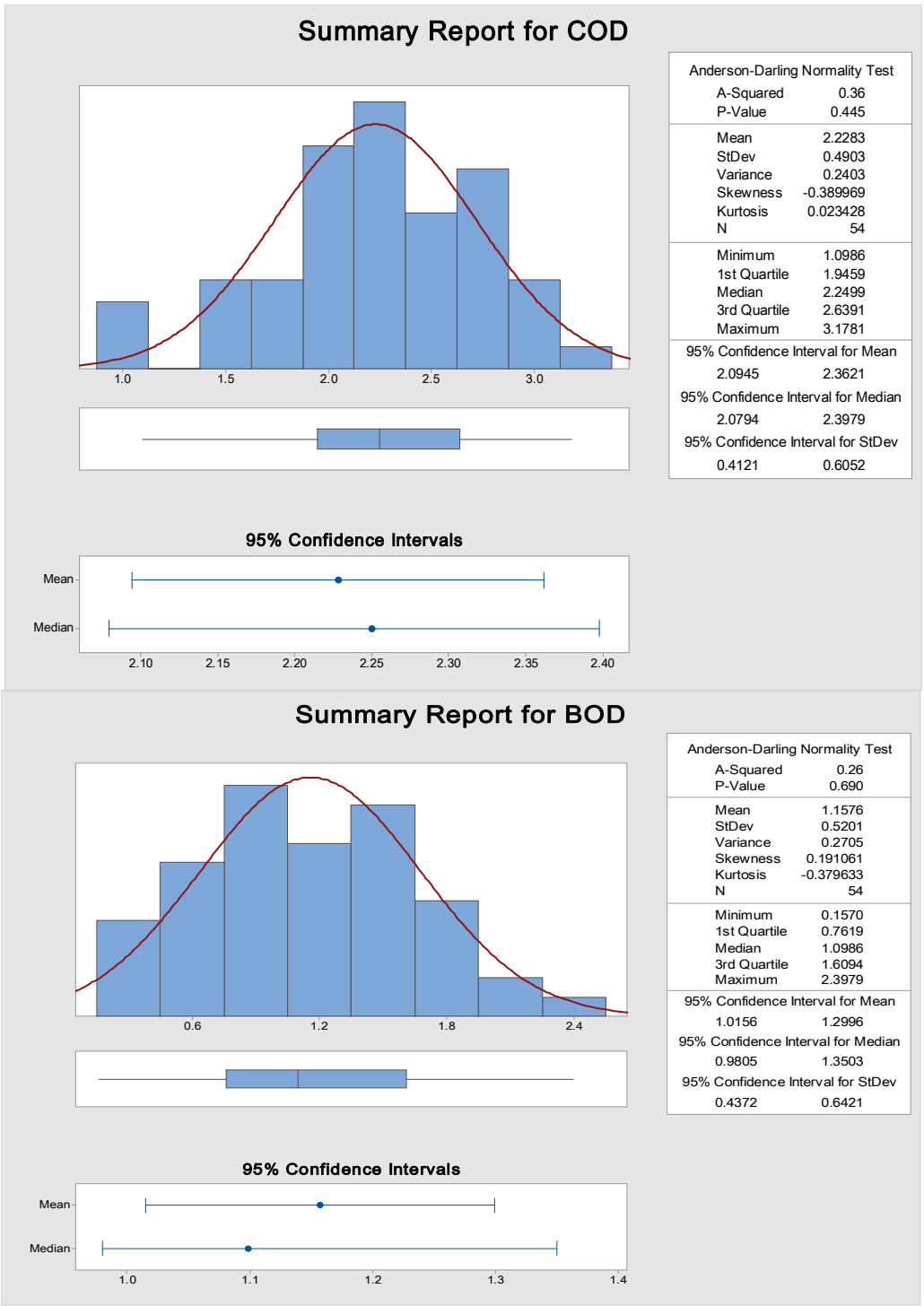


Figure (4): Histogram, boxplot and confidence intervals of some variables (ln values) along the Nile during August (2010-2012)

Figure 5 presents the eigenvalues for the set of variates in the low flow condition (February). The eigenvalue is the variance explained by a variate or a factor. The individual eigenvalues indicate the captured variance, the higher value interpreted to more captured variance. Factor-1(F-1), Factor-2(F-2), Factor-3(F-3) and Factor-4(F-4) with values of eigenvalues about one or greater, explain most of the data set. As listed in table 3, these factors represents 25.6%, 17.2%, 17.1% and 14.8%, respectively, of the total variance of the data. By classifying the component loading according to **Liu *et al.* (2003)**, the

loading values which are greater than 0.75 signify indicate as "strong", the loading with absolute values range between 0.50 and 0.75 indicate as "moderate" while loading values range between 0.30 and 0.50 indicate as "weak". Based on these classifications, the first factor is highly affected by TDS (loading 0.90) and moderately affected by nitrate and TSS (loading <0.75) which represents the agricultural activities. The second factor is highly affected by pH and moderately affected by DO. Negative factor loading of OP indicates the disproportion between this parameter and F-2. pH and DO which were correlated with F-2, increased with decreasing OP concentration which caused by photosynthesis activity in water and basically reflects biological activity of the Nile. The third factor (F-3) is mainly made by high loading of Fecal Coliform (FC) and Total Coliform (TC) and represents the bacterial group pollutants which points to some source of domestic wastewater discharges and could be identified as "bacterial contamination". The fourth factor (F-4) is highly affected by COD and BOD which represents the organic pollutant and reflects the domestic wastewater discharge, as shown in figure 6.

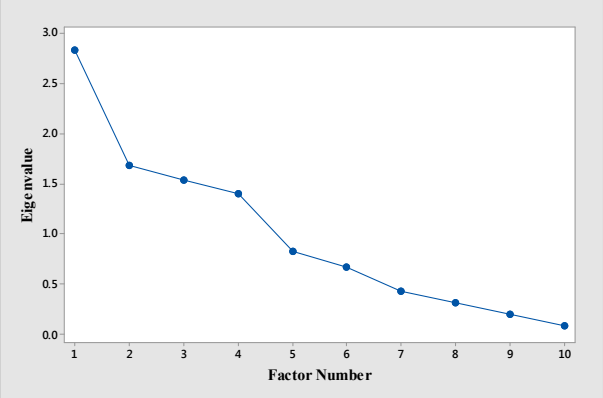


Figure (5): Scree plot of the eigenvalues

Table (3): Varimax rotated factors for the first four rotated variates along the Nile during February

Variable	Factor Loading			
	Factor-1 (F-1)	Factor-2 (F-2)	Factor-3 (F-3)	Factor-4 (F-4)
NO_3	**-0.676**	0.231	-0.079	-0.063
Ortho-P	0.514	**-0.694**	-0.128	-0.052
COD	0.385	0.115	-0.091	**-0.878**
BOD	-0.465	-0.173	0.052	**-0.804**
TSS	**-0.725**	-0.322	0.167	0.003
FC	0.010	-0.013	**-0.895**	-0.084
TC	0.123	0.025	**-0.900**	0.052
pH	0.066	**0.792**	0.023	-0.100
DO	0.351	**0.622**	-0.150	0.180
TDS	**-0.900**	-0.144	0.125	0.091
Eigen value	2.838	1.687	1.538	1.408
% Variance	25.60%	17.20%	17.10%	14.8%
% Cumulative	25.60%	42.80%	59.9.90%	74.7%

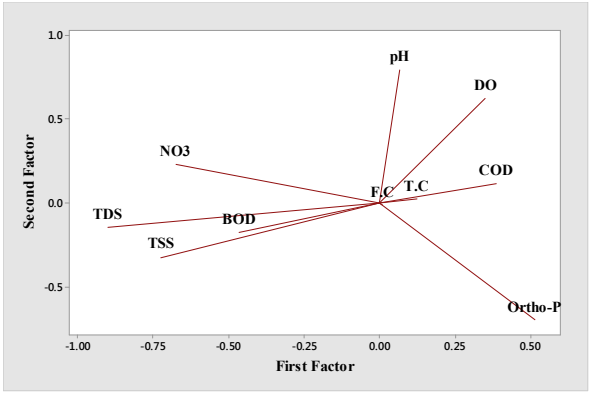


Figure (6): Varimax rotated factors along the Nile during February

Table 4 represents the correlation components matrix (component score covariance matrix) of varimax rotated four Principal Component (PC) which indicate that there is no correlation between components so each component represent an discrete unit from others.

Table (4): Component Score Covariance Matrix

Component	1	2	3	4
1	1.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000
3	0.000	0.000	1.000	0.000
4	0.000	0.000	0.000	1.000

Table (5): Varimax rotated factors for the first three rotated variates along the Nile during August

Variable	Factor Loading		
	Factor-1 (F-1)	Factor-2 (F-2)	Factor-3 (F-3)
NO_3	**0.937**	0.091	-0.082
Ortho-P	-0.331	-0.247	0.243
COD	-0.141	-0.512	-0.009
BOD	**0.761**	-0.255	0.090
TSS	0.544	0.081	**-0.638**
FC	-0.049	**-0.945**	0.037
TC	0.060	**-0.903**	0.095
pH	-0.085	0.026	**-0.938**
DO	0.330	0.085	**-0.558**
TDS	**0.847**	0.042	-0.431
Eigenvalue	3.489	2.311	1.262
% Variance	27.20%	21.20%	18.70%
% Cumulative	27.20%	48.40%	67.10%

In high flow along the Nile, August, the eigenvalues for the set of variates are presented by figure 7. As listed in table 5, the first three factors represents ratios of 27.2%, 21.2%, and 18.7%, respectively, of the total variance of the data with total ratio 67.1% of the total variability of the original data. Performing Varimax normalized rotation gives three variates easier to interpret as shown in table 5. The first factor (F-1) is highly affected by TDS, BOD, and nitrate, which reflects the mixing of

organic contamination in agriculture. The second factor (F-2) gives information about the variation in the level of FC and TC, and represents the bacterial pollutant and reflects the domestic wastewater discharge. The third factor (F-3) is highly affected by pH and moderately affected by TSS and DO and reflects the biological activity in water, as shown in figure 8.

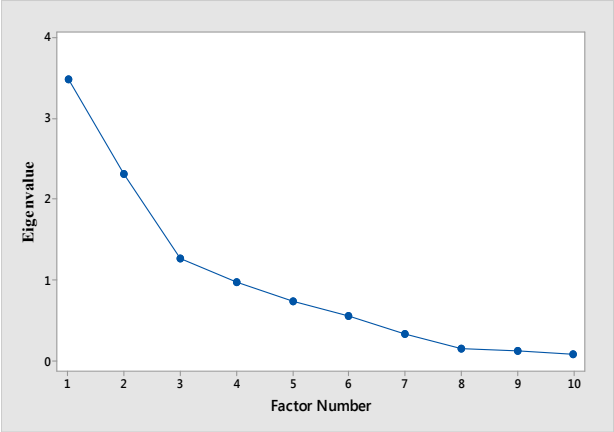


Figure (7): Scree plot of the eigenvalues

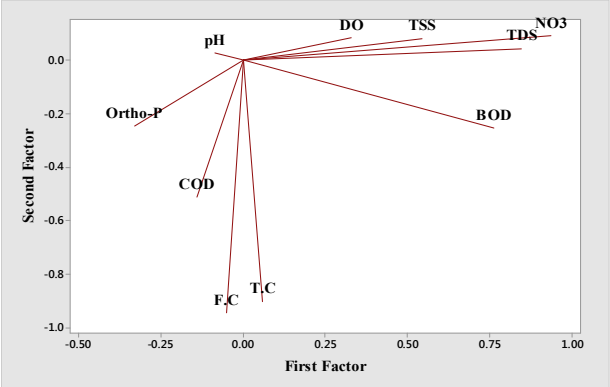


Figure (8): Varimax rotated factors along the Nile during August

Table 6 represents the correlation components matrix (component score covariance matrix) of varimax rotated three Principal Component (PC) which indicate that there is no correlation between components so each component represent an discrete unit from others.

Table (6): Component Score Covariance Matrix

Component	1	2	3
1	1.000	0.000	0.000
2	0.000	1.000	0.000
3	0.000	0.000	1.000

In general, the three extracted factors represent three different processing of agricultural pollutants, biological activity and bacterial pollutants.

A group of water quality parameters are used in the water quality index calculations along the Nile. Factor analysis used to extract the most important indicator parameters used in the calculation of water quality index during both flow conditions as listed in table 7. It can be concluded that the water quality index can rely on the water quality parameters TDS, TSS, BOD, NO3, pH, DO, FC and TC.

Table (7): Most important variables contributing to each of the first three rotated variates

Water body	F-1	F-2	F-3
Nile (February)	TDS+TSS+ NO_3	pH+ DO+OP	FC+TC
Nile (August)	TDS+BOD+NO_3	FC+TC	pH+DO+TSS

Conclusions and Recommendations

Comparing the mean value of each variable with the national standards of Law 48/year 1982, all water quality parameters have mean values within the allowable limits except COD, which recorded values slightly above the allowable limits. Factor Analysis revealed that three and four factors are sufficient to explain the data variability. These factors explain more than 74.7% and 67.1% of the total variance of the collected data set along the Nile during February and August, respectively. The Varimax rotated variate loadings indicate that variables responsible for water quality variations are mainly related to agricultural uses (TDS, TSS, and nitrate) during February and agricultural uses mixed with domestic wastewater (TDS, BOD, and nitrate) during August. The bacterial pollutant represents the second and third factors during August and February, respectively. The most important indicator parameters used in water quality index calculation were found TDS, TSS, NO_3, BOD, FC, TC, pH and DO. Factor analysis provides a useful tool that could help the decision makers in determining the extent of pollution via practical pollution indicators. It could also provide a crude guideline for selecting the priorities of possible prevention measures in the proper management of the surface water resources of the Nile. It is recommended to use the factor analysis as a tool for identifying the most important water quality parameters representing the main sources of pollution not only in the River Nile but also in irrigation canals and agricultural drains.

References

1. APHA., 1998. Standard Methods for the Examination of Water and Wastewater." Washington, DC.
2. Canadian Council of Ministers of the Environment (CCME), (2001). Canadian water quality guidelines for the protection of aquatic life, CCME water quality Index 1, 0". User`s manual. Excerpt from Publication No.1299, ISBN 1-896997-34-1.
3. Gupta AK, Gupta SK and Patil RS (2005). "Statistical analyses of coastal water quality for a port and harbour region in India. Environ. Monit". Asses. 102 179-200.
4. Kowalkowski, T., R. Zbytniewski, J. Szpejna and B. Buszewski, 2006. "Application of Chemometrics in River Water Classification". Journal of Water Research, 40(4): 744-752.
5. Liu CW, Lin K H, Kuo YM (2003). Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. Sci. Total Environ. 313:77-89.
6. Nouri, J.; Karbassi, A. R.; Mirkia, S., (2008). "Environmental management of coastal regions in the Caspian Sea. Int. J. Environ". Sci. Tech., 5 (1), 43-52.
7. Praus. P.; (2005). "Water quality assessment using SVD-based principal component analysis of hydrological data".
8. Sargaonkar, A. and Deshpande, V., (2003). "Development of an overall index of pollution for surface water based on a general classification scheme in Indian context". Environ. Monit. Assess 89 43-67.
9. Shrestha, S.; Kazama, F., (2007). "Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan". Environ. Model. Software, 22 (4), 464-475.
10. Singh, K. P.; Malik, A Mohan, D.; Sinha, S., (2004). "Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): A case study". Water Res., 38 (18), 3980-3992).
11. Singh, K. P.; Malik, A Mohan, D.; Sinha, S., (2005). "Water quality assessment and apportionment of pollution sources of Gomti River (India) using multivariate statistical techniques: A case study". Analytica. Chimica. Acta, 538 (1-2), 355-374.
12. Zeng, X. and Rasmussen, T.C., (2005). "Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA. J. Environ". Qual. 34 1980-1991.

1/23/2015