Asian Journal of Water, Environment and Pollution, Vol. 9, No. 4 (2012), pp. 1–10.

Application of Multivariate Statistical Analysis to Define Water Quality in Jajrud River

G. Asadollahfardi*, A. Kodadadi[1], B. Paykani[2], Y. Samady[3] and R. Asadollahfardi[4]

Civil Engineering Department, Kharazmi University, Tehran
[1]Faculty of Engineering, Tarbiat Modares University, Tehran
[2]Environmental Engineering Expert, Tarbiat Moallem University, Tehran
[3]Faculty of Computer and Statistics, Shahid Beheshdi University, Tehran
[4]Water and Environment Consultant, Vancouver, Canada
✉ asadollahfardi@yahoo.com

Received September 14, 2011; revised and accepted September 14, 2012

Abstract: The copious prevalence of water deficiency and the geographical location of Iran (arid and semi-arid zone) make acquiring enough accurate data of water quantity and quality for water management vital. However, merely having sufficient data without proper interpretation is rather worthless too when it comes to effective water management and thus, there are several techniques for analyzing water quantity and quality. In this work, statistical method was used to analyze the data collected from the catchment area under study i.e. Jajrud River, located in the North West of Tehran Province. The multivariate time series method was employed to analyze water quality parameters in the river. Box-Jenkins time series model was also applied to the factor data resulted from the Multivariate time series. The results showed that the water quality parameters are not independent having a correlation coefficient larger than 0.3. The study also shows that ground water is the first effective factor, which causes increasing total dissolved solid (TDS) in the river. Domestic waste water pollution is the second-most important factor. Agricultural fertilizers and industrial waste may rank as the third and fourth pollution factors, respectively. Prediction of factor data using Box-Jenkins model was accurate and suitable which may be applicable to other place to model the factors data instead of many water quality parameters.

Key words: Water quality, Jajrud river, multivariate statistical technique, principal analysis, Box-Jenkins model.

Introduction

Every water body has a certain capacity of self-purification and within that capacity, it is perfectly capable of cleansing itself. However, increasing population, consumption and land usage are causing increased water pollution and thus the amount of polluted water discharge in water bodies are regularly exceeding their self-purification capacities and therefore, the natural purification process cannot treat the water entirely. Given the lack of water in Iran, identification of water resource, non-point and point pollution sources and prediction of water quality trend for decision makers are of critical importance. Jajrud River, by virtue of its location near Tehran, capital of Iran, supplies part of Tehran's drinking water. Hence, the protection of water quality in the river is important. The runoff from the surrounding regions washes the catchment area, mingling domestic and agricultural pollution in the water. Figure 1 shows the situation of catchment area on the map of Iran.

Vega et al. (1996) employed an ANOVA of rotated principal component to physico-chemical water quality

*Corresponding Author

Abbreviations

MSE	=	Mean Square Error
SAS	=	Statistical Analysis System
SSE	=	Sum of Square Error
MAE	=	Mean Absolute Error
RS	=	R-Square
ARS	=	Adjusted R-Square
AIC	=	Akaike Information Criterion
SBIC	=	Schwarz Bayesian Information Criterion
Chisq	=	The chi-square statistic
Df	=	Degree Freedom
Pval	=	P Value
RMSEA	=	Root Mean Square Error of Approximation
CFI	=	Comparative Fit Index
GFI	=	Goodness Fit Index
SPSS	=	Statistical Package for Social Science
ARMA	=	Autoregressive Moving Average
SARIMA	=	Seasonal Autoregressive Integrated Moving Average
SAF	=	Sample Autocorrelation Function
PAF	=	Partial Autocorrelation Function

parameters. The results showed that the mineral content of the water are seasonal and climate dependent, thus pointing to a natural origin for this polluting form and secondly that pollution by organic material originates from anthropogenic sources, mainly as municipal wastewater. Qiming et al. (1996) analyzed water quality in Meilirng bay and part of western Taihu Lake (China) using principal component and found that there exists an obvious spatial and temporal variation in the main factor of water quality. Mazlum et al. (1999) studied the water quality variation factor in Agackoy monitoring station on the Porsuk tributary in the Sakarya River basin using principal component. The result demonstrated that the method can be reliably used for water quality analysis. Perona et al. (1999) studied the spatial and temporal variation in water quality of a Spanish river (Alberche river) with two years data using principal component analysis. The results showed that the variation in nutrient content in the river can be attributed to anthropogenic sources since in this season, the presence of holiday makers leads to high population density in the nearby residential building and recreational areas.

Alberto et al. (2001) employed multivariate statistical techniques for water quality in the Suquia River (Cordoba-Argentina). The discriminant analysis showed the best results for data reduction and pattern recognition during both spatial analyses. Simeonv et al. (2003) worked on water quality with data collected for three years from major river system (Aiakmon, Axios, Gallikos, Loudias and Strymon) in northern Greece using

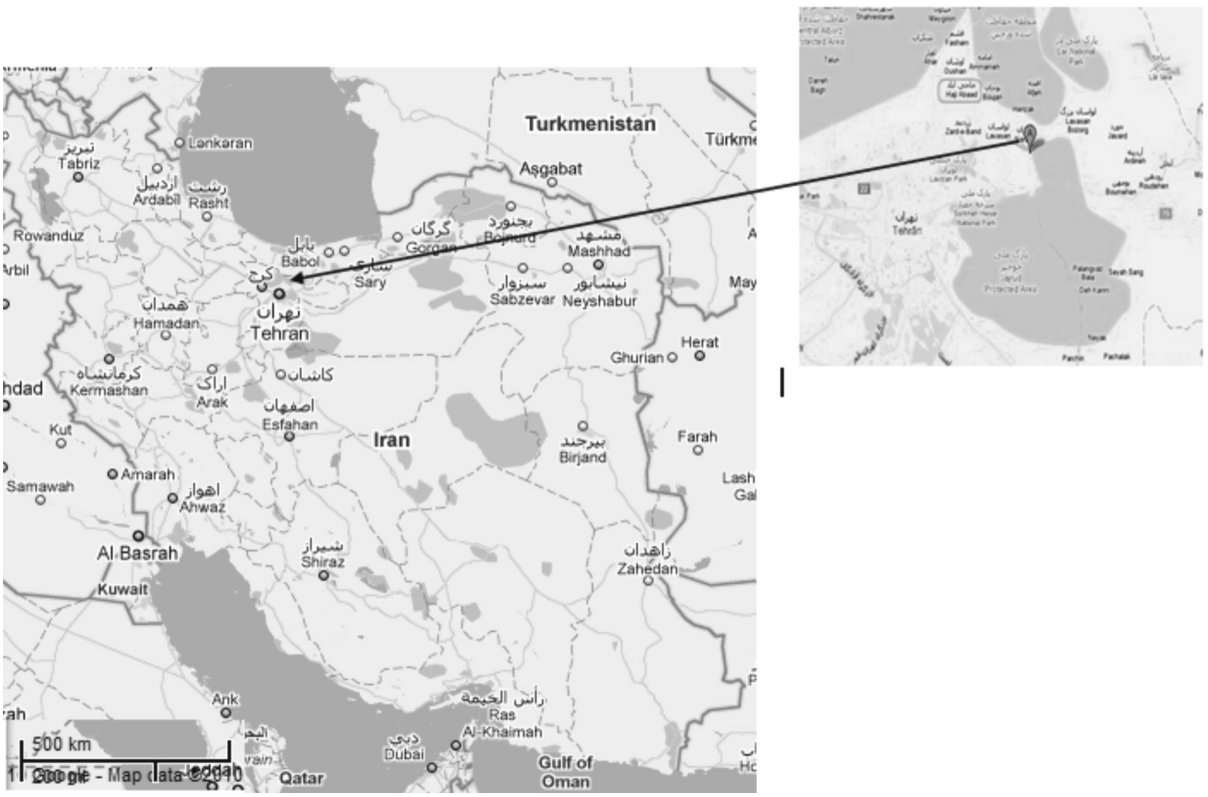

Figure 1: Study area.

different multivariate statistical methods. The results showed that the multivariate statistical methods are proper for surface water analysis. Bengrane and Marhaba (2003) used principal component of chemical, biological and physical water quality data which was monitored at 12 locations along the Passaic River, New Jersey, during the year 1998. The result showed that principal component has a proper method for analyzing water quality. Parinet et al. (2004) used principal component analysis using coefficient of linear regression to describe the tropic state of an eutrophic lake system in Ivory Coast. The results also led evidence to the suitability of this method. Singh et al. (2005) used multivariate statistical method to water quality data of the Gomati River (India) which was gathered for 34 different parameters at different stations during a period of three years. The results demonstrated the usefulness of multivariate statistical methods for interpreting a large number of water quality data.

Zeng and Rasmusson (2005) used multivariate statistical method for reducing the number of measurements of parameters, locations and frequency without compromising the quality of the monitoring programme, and the results were satisfactory. Ouyang et al. (2006) studied water quality data for 16 physical and chemical parameter collected at 22 monitoring stations in a river using principal component. The results indicated that the parameter that is most important for a season may not hold the same importance in another season except for DOC and electrical conductivity. Shrestha and Kazama (2007) applied multivariate statistical techniques for the evaluation of temporal/spatial variations and the interpretation of a large complex data set collected from the Fuji River basin during eight years monitoring period of 12 water quality parameters at 13 different sites. The results were satisfactory. Iscen et al. (2008) studied the water quality of Uluabat Lake (Turkey) using a different multivariate statistical approach; the result showed the microbiological factor explained 32.34% of total variance, while a second factor, the organic nutrient factor accounted for 25.46% and the third factor, the physicochemical factor explained 19.54% of the variance. Kazi et al. (2009) employed multivariate statistical techniques for water quality analysis of Machar Lake (Pakistan) with data amassed for 35 water quality parameters during 2005-2006 at five monitoring stations. The results revealed that the major causes of water quality deterioration were related to the inflow of industrial, domestic and agricultural effluents as well as seepage of saline into the lake at a site. Fishing and boating at two other sites also contributed to diminution of water quality in the lake.

Some researchers applied Box-Jenkins time series of water quality parameters such as Huck and Farquhar (1974), Lohani and Wang (1987), Jayawardena and Lai (1987), Asadollahfardi (2003) and Renwick et al. (2006); however, none of the them applied the mentioned method to factor data.

Jajrud River is situated between longitude 51° 25′ and 51° 55′ and between latitude 35° 45′ and 36° in North East of Tehran province (Iran). The basin is endowed with mild weather during summer and cold in winter, so people of Tehran city travel to the area during the weekend especially in summer. The length of the river is about 140 kilometres with a slope of 2.2 percentage. The river is one of the main drinking water resources for the people of Tehran. The mountains from which the river originates consist of calcite, dolomite, silt and Marn clay. Latian Dam has been constructed above the river and has a catchment area about 690 square kilometres. Figure 1 shows the situation of the Jajrud River in Iran.

The first objective of this work is to apply multivariate statistical method to water quality data containing electrical conductivity (EC), total dissolved solid (TDS), ammonia (NH_3), nitrite (NO^-_2), nitrate (NO^-_3), chemical oxygen demand (COD), biochemical oxygen demand (BOD_5), total coli form (TC) and fical coli form (FC) data which were collected from Water Authorities of Tehran. The second aim is to develop Box-Jenkins time series model using the factor data which is resulted from first objective.

Materials and Methods

Theory

Multivariate analysis involves analysis of several variables simultaneously and as such an intricate task. One of the methods used to solve this type of problem is known as "Factor analysis". In this method, parameters which have a high correlation coefficient between them are gathered into one single group, which is then called a Factor. Factor analysis is used to understand the correct structure of collected data and to identify the most important factors contributing to the data structure (Schoer, 1985; Buckley and Winters, 1992; Padro et al., 1993). The Factor is a new independent variable, which cannot be directly measured and is in fact, immeasurable. Factor analysis is also applied to find associations between parameters so that the number of parameters measured can be reduced. Equation (1) describes the factor (Johnson and Wichem, 2007).

$$F_j = \sum_{i=1}^{p} w_{ji} x_i = w_{j_1} x_1 + w_{j_2} x_2 + \ldots + w_{jp} x_p \qquad (1)$$

where F_j is factor, W_{ji} is factor number, 1 … p is number of variables and x_i is observed variables.

Procedure of Factor Analysis Implementation

Implementation of factor analysis involves the following steps:

1. Gathering all of the data, and creation of a correlation matrix of all variables (parameters) used in the analysis
2. Finding the factors loading according to correlation coefficients.
3. Rotating the factors for simplicity and understandability of factor analysis.

Box-Jenkins Methodology for Time Series Modelling

Decomposition of time series data into its components, while being instructive and revealing, is a difficult job. Moreover, it causes greater errors by accumulation of component errors. To avoid these difficulties, Box and Jenkins (1976) developed a new methodology which, in essence, does the same job but unifies all concepts discussed above. In this method, using some transformations such as simple and seasonal differences, the trends, seasonal and cyclical components present in the data are removed. Then, a family of models is entertained for the transformed data, which is expected to be as simple as possible.

The Box-Jenkins approach is based on the notion of stationary time series briefly explained in the following section.

Classification of Non-seasonal Time Series Models

The general non-seasonal autoregressive moving average model of order (p, q) is (Eq. 2):

$$Z_t = \delta + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (2)$$

where $\phi_p(B)$ and $\theta_q(B)$ are the autoregressive and moving average operators, respectively, defined as:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B_p \qquad (3)$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B_q \qquad (4)$$

where B is the backward shift operator, so that

$$B^k Z_t = Z_{t-k}.$$

When a series of $Z_t, Z_{t-1}, \ldots, Z_n$ show nonstationarity, i.e., the mean and variance of the series are changing with t, it may still be related to the random deviates a_t by means of the following model.

$$\phi_p(B)\nabla^d Z_t = \theta_q(B) a_t \qquad (5)$$

where ∇^d equals the backward difference operator. Equation (5) represents the Autoregressive Integrated Moving Average (ARIMA) (p, d, q) model with integers p, d and q defining the order of the model. Essentially, the Box-Jenkins procedure consists of four basic steps, which are shown in Table 1 and Figure 2. For more detail on the Box-Jenkins model structure and forecasting, refer to Box and Jenkins (1976).

Table 1: Stages of Box-Jenkins modelling

	Step Description
1.	**Check the data for normality**
	(a) No transformation
	(b) Square root transformation
	(c) Logarithmic transformation
	(d) Power transformation
2.	**Identification**
	(a) Plot of the transformed series
	(b) Autocorrelation function (ACF)
	(c) Partial autocorrelation function (PACF)
3.	**Estimation**
	(a) Maximum likelihood estimate (MLE) for the model parameters
	(Ansley algorithm)
4.	**Diagnostic checks**
	(a) Overfitting
	(b) Examination of the residuals (modified Portmanteau test)
5.	**Model structure selection criteria**
	(a) AIC criteria
	(b) PP criteria
	(c) BIC criteria

Data Analysis

Three years worth of data from March 2006 to March 2009, monitored by the Tehran Regional Water Authority, was used for this study. The parameters studied were EC, TDS, NH_3, NO_2, NO_3, TC, FC, BOD_5 and COD. Table 2 shows a statistical summary of the data.

All monitored water quality listed in Table 2 is employed in the multivariate statistical analysis for the purpose of testing whether the parameters can be used in factor analysis. Sampling dates with missing observations were excluded. Merely, the mean monthly value was used. The multivariate analysis was conducted using SAS software (SAS, 1993).

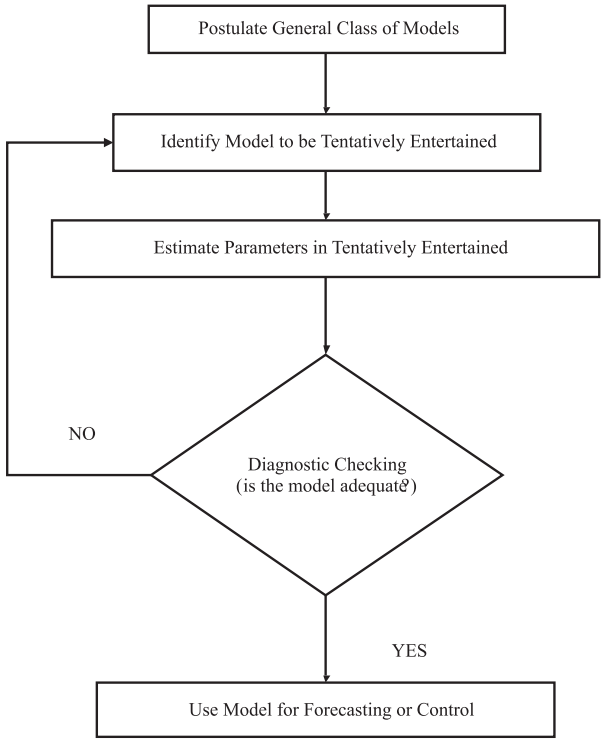

Figure 2: Stages of the iterative approach to model building.

Table 2: Statistical summary of Jajrud River water quality data

Definition	*EC (mμ/s)*	*TDS (mg/l)*	*NH_3 (mg/l)*	*NO_2 (mg/l)*	*NO_3 (mg/l)*	*COD (mg/l)*	*BOD_5 (mg/l)*	*TC (No./100 ml)*	*FC (No./100 ml)*
Skeness	-0.33	0.02	1.34	0.20	0.54	0.06	1.24	1.18	0.09
Coeff Variation	7.61	8.15	28.12	15. 97	15.43	25.27	12.44	75.55	45.31
Kurtosis	-0.24	-0.91	1.65	-0.58	-0.26	-0.91	1.39	0.39	-0.098
Std Error Mean	4.88	3.05	0.00	0.00	0.09	0.13	0.03	8065.30	742.11
Mean	384.60	224.43	0.10	0/04	3.50	2.99	1.35	64050.42	9826.11
Median	391.09	224.80	0.09	0/04	3.46	3.08	1.30	40907	9750
Mode	326.19	211.35	0.09	0/04	3.47	2.47	1.28	-	-
Std Deviation	29.28	18.30	0.03	0.01	0.54	0.76	0.17	48392	4453
Variance	857.08	334.78	0.0008	0.000034	0.29	0.57	0.03	2341765935	19826290
Range	115.30	69.12	0.12	0/02	2.23	2.78	0.71	173970	15678
Q_3 - Q_1	36.17	27.71	0.04	0/01	0.68	1.13	0.19	55210	6520
Percentile 100%Max	441.49	263.70	0.19	0/05	4.33	1.84	191451	18149	18149
Percentile 99%	441.49	263.70	0.19	0.05	4.33	1.84	191451	18149	18149
Percentile 95%	441.49	249.50	0.17	0.05	4.23	1.73	172569	17637	17637
Percentile 90%	413.90	246.21	0.14	0.04	4.21	1.55	140568	16287	16287
Percentile 75% Q_3	401.52	239.06	0.12	0.04	3.50	1.42	82757	12861	12861
Percentile 50% Median	391.09	224.80	0.09	0.04	3.08	1.30	40907	9750	9750
Percentile 25% Q_1	365.35	211.35	0.08	0.03	2.37	1.23	27547	6341	6341
Percentile 10%	336.79	198.23	0.07	0.03	2.03	1.17	19456	3564	3564
Percentile 5%	326.19	197.20	0.07	0.03	1.84	1.16	18754	3250	3250
Percentile 1%	326.19	194.58	0.07	0.03	1.55	1.13	17481	2471	2471
Percentile 0% Min	326.19	194.58	0.07	0.03	1.55	1.13	17481	2471	2471

First of all, the correlation matrix shown in Table 3 was constructed. As shown in the table, the numbers of the correlation coefficients are above 0.3, which suggests that factor analysis can be used (Tabachnick and Fidell, 2001).

Table 3 shows the correlation matrix for observed water quality parameters data in the monitoring station. Some of the parameters are positively correlated to each other, including the correlation between EC and TDS, $r = 0.742$, TC and EC, $r = 0.564$, TC and TDS, $r = 0.476$, nitrite (NO_2) and COD, $r = 0.416$ and BOD_5 and NO_3, $r = 0.392$. Some other water quality parameters are negatively correlated notably COD and NH_3, $r = –0.344$. Correlation between COD and EC is $r = –0.327$. Other parameters have much smaller correlations coefficients, which negates the need of considering them and suggests that they may affect the water quality independent of each other.

One of the techniques used to figure out the number of factors is the estimation of eigenvalue and factor loadings for the correlation matrix, and each eigenvalue corresponded to an eigenvector that identifies the group of water quality parameters that were most highly correlated among them. The first eignfactor accounted for greatest variation among the observed water quality parameters, while each following eigenfactors was orthogonal to all former factors, and provided incrementally smaller contributions to the overall descriptive ability of the model. The eigenvalue of the correlation matrix of this study is shown in Table 3. Since, a lower eigenvalue may contribute only modestly to the descriptive ability of the water quality data, merely the first few factors were chosen. Methods are present to figure out the number of factors that need to be considered and the number of those that can be safely ignored (Browne, 1968; Linn, 1968; Tucker et al., 1969; Hakstian et al., 1982). The method of Kaiser criterion, which retains merely those factors with eigenvalue bigger than one, is most widely used technique (Kaiser, 1960).

Table 3 provides the eigenvalue and explanatory capability for water quality data. The first factor merely accounts for 45% of total variability, while following factors provide a diminishing ability to predict water quality variations. First four factors each has eigenvalue bigger than one, and the four factors shows 74% of water quality variables..

Table 4 shows individual and cumulative eigenvalue of the river water quality observations. Also shown are individual and cumulative contributions factors toward explaining water quality variation.

Table 3: Correlation coefficient matrix between the water quality parameters

	EC	*TDS*	*NH_3*	*NO_2*	*NO_3*	*TC*	*FC*	*BOD_5*	*COD*
EC	1	0.742	0.08	-0.047	-0.463	0.564	0.343	0.021	-0.327
TDS	0.742	1	0.02	-0.113	-0.442	0.476	0.153	-0.178	-0.239
NH_3	0.08	0.020	1	-0.386	0.162	-0.372	-0.096	0.289	-0.344
NO_2	-0.047	-0.113	-0.386	1	0.198	0.205	0.340	0.189	0.416
NO_3	-0.463	-0.442	0.162	0.198	1	-0.313	-0.102	0.392	0.132
TC	0.564	0.476	-0.372	0.205	-0.313	1	0.120	-0.061	-0.184
FC	0.343	0.153	-0.096	0.340	-0.102	0.120	1	0.321	0.125
BOD_5	0.021	-0.178	0.289	0.189	0.392	-0.061	0.321	1	-0.190
COD	-0.327	-0.239	-0.344	0.416	0.132	-0.184	0.125	-0.190	1

Table 4: Individual and cumulative eigenvalue of the river water quality observations

Factor	*Eigenvalue*		*Variance*	
	Individual	*Cumulative*	*Individual*	*Cumulative*
1	2.1348	2.1348	0.24	0.24
2	1.9054	4.0417	0.210	0.45
3	1.4284	5.4685	0.160	0.61
4	1.1670	6.6355	0.130	0.74
5	0.8164	7.4519	0.090	0.83
6	0.5378	7.9898	0.060	0.89
7	0.3811	8.3709	0.043	0.93
8	0.3325	8.7034	0.037	0.97
9	0.2966	9.0000	0.030	1.00

As shown in Table 5, the *p*-value for zero factor model, one factor and two factors are relatively small, so the models with less than three factors are not appropriate, but the *p*-value for four and five factors are above 0.5 which can thus be selected as a candidate. Usually, if RMSE is less than 0.5, the model is well fitted, and if the RMSE is less than 0.1, the model is approximately good fitted (Johnson and Wichern, 2007). According to this definition, four-factor models is suitable (Table 5). If the amount of R^2 approaches to one, the model is well fitted. By this definition, the four factors and five factors are appropriate (Table 5). Another index which can be applied is AIC; according to this technique, if the amount of AIC is small, this meant that the model is suitable. According to this method, four factors are proper. In the second part of Table 4, there is comparison with consecutive factors. As shown in Table 3 the *p*-value preference of three factor related to four factor are rejected with $a = 0.05 > p$-value = 0.049. This means that four factor is preferred. Since the assumption of preference of the four factor as compared to five factor could not be rejected ($a<p$-value, 0.5<0.53), four factors are more suitable than five factor.

Factor loading reflects the correlation between the water quality parameters and the extracted factors. Factor loadings for the four retained eigenvalues are shown in Table 6. Factor loading is shown with rotation using Quartimax method. The main function of the factor rotation application is to facilitate interpretation by providing a simple factor structure. The factors were rotated in order that the observed axes were aligned with a dominant set of water quality parameter which assisted in understanding the relation of factors to the observed water quality parameter (Zeng and Rasmusson, 2005). In this work the Quartimax rotation was applied, another rotation such as biquartimax, equamax and varimax were also developed (Johnson and Wichern, 2007; Kaiser, 1958). Factor loading >0.5 is considered significant in this work. The first factor incorporates those water quality parameters, which may be characteristic of ground water including EC and TDS because ground water can enter into the river increasing the amount of TDS. Since the water of the river is used for drinking, measures are in place to protect the water body and inhibit the inflow of waste water. The mean value of ammonia, nitrite and

Table 5: The suitable statistical summary for some numbers of factors

	0	*1*	*2*	*3*	*4*	*5*	*Saturated*
Chisq	76.66	53.57	32.14	16.72	4.11	0.03	0
Df	36	27	19	12	6	1	0
Pval	9.19×10^{-5}	17.15×10^{-4}	0.03	0.16	0.66	0087	NA
RMSEA	0.180	0.17	0.14	0.11	0	0	NA
CFI	0.00	0.35	0.68	0.88	1	1	1
GFI	0.70	0.75	0.84	0.91	0.97	1	1
AGFI	0.62	0.59	0.63	0.65	0.81	0.99	1
AIC	4.66	-0.43	-5.86	-7.28	-7.89	-1.97	0

	0 vs 1	*1 vs 2*	*2 vs 3*	*3 vs 4*	*4 vs 5*	*5 is saturated*
Chisq	23.084	21.431	15.423	12.607	4.085	0.026
Df	9	8	7	6	5	1
Pval	0.0060	0.0061	0.0309	0.0497	0.5372	0.8710

Table 6: Factor loading (rotated) for the water quality observations

Parameter	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*	*Communality*
EC	0.9137	0.1536	-0.0424	0.2776	0.995
TDS	0.4791	0.0272	0.1204	-0.203	0.3009
NH_3	0.0117	-0.1921	0.3016	0.4121	0.2797
NO_2	-0.0055	-0.1239	-0.7408	-0.0368	0.5672
NO_3	-0.1756	0.0486	-0.5904	0.2299	0.4845
TC	0.1023	-0.9563	-0.1580	-0.1385	0.9950
FC	0.2521	-0.5222	-0.2589	-0/1522	0/3383
BOD_5	-0/0973	0/2533	-0.1769	0.9531	0.9950
COD	-0.0493	-0.1226	-0.2384	0.4166	0.3561

BOD perhaps prove the effectiveness of these measures (Table 2).

The second factor has relatively higher factor as compared to the first factor, with both being negative. The largest negative value –0.956 belongs to TC, and second negative value is –0.522. TC can originate from land, air and domestic wastewater, but FC can be derived only from domestic wastewater. The high density of tourists that travel to the regions especially in late Spring, Summer and early Fall greatly add pollutants to the river. Also, people living in the mountain area discharge waste water to the river without proper treatment.

The third factor has a little lower factor loading, the largest negative value (–0.74079) being for NO_2, and second negative value (–0.59041) belonging to NO_3. As shown in Table 2, mean value of NO_2 and NO_3 water quality parameters is satisfactory for drinking water (WHO, 2008). The concentration of NO_2 and NO_3 may spring from the fertilizers used by farmers, or alternatively, it could also originate from waste water added by living people in the mountainous area of the basin. The fourth factor is strongly related to BOD (r = 0.953), and NH_3 and COD have a lower factor loading which is 0.412 and 0.41665. The fourth factor loading may show some industrial pollution but if considering mean value of the water quality parameters in Table 2, the factors certainly have less priority than others.

Model Developing

According to mentioned Box-Jenkins methodology and using SAS software, we developed an ARIMA model for factor 1. Different ARIMA models were obtained with different statistical testing which are shown in Table 5. According to the result of the modelling (Table 7), ARIMA p = (1,3,4) d = (1,12) , q = (1,3,4) is suitable model, because its R^2 and adjust R^2 are bigger than other models, and other statistical parameters are less.

Hence, the model can be written as following equation

ARIMA p = (1,3,4) d = (1,12) q = (1,3,4)

According to mentioned methodology, it was computed parameters of the model. The parameters are shown in Table 8.

Hence, the model and parameters can be shown in following equations

$$\phi_p(B)\,\phi_p(B^L)\,\nabla_L^D\,\nabla^d y_t^* = \theta_q(B)\,\theta_Q(B^L)\,a_t$$

where $\theta_Q(B^L) = 1$, $\phi_p(B^L) = 1$

Table 7: The amount of statistical testing for different ARIMA models

Model	*SSE*	*MSE*	*MAE*	R^2	*AJD* R^2	*AIC*	*SBC*
ARIMA (0,1,1) s	9.7811	0.4075	0.4491	0.678	0.630	-18.5423	-17.3640
ARIMA (0,1,0) (0,1,1) s	12.2807	0.5339	0.5217	0.629	0.570	-12.4316	-11.2961
ARIMA (0,0,1) (0,1,1) s	8.6029	0.3584	0.4482	0.743	0.731	-18.6229	-17.2667
ARIMA (0,1,1) (1,0,0) s	25.2782	0.7222	0.6553	0.477	0.461	-7.38919	-4.2785
ARIMA (2,0,0) (1,0,0) s	20.5679	0.57133	0.5159	0.631	0.597	-12.1522	-5.8181
ARIMA (4,1,0) (0,1,0) s	7.4628	0.3245	0.4041	0.720	0.675	-17.8880	-13.3460
ARIMA (0,1,2) (0,1,1) s	9/0447	0.3932	0.4833	0.660	0.626	-15.4661	-12.0596
ARIMA (0,1,2) (0,1,0) s	9.0458	0.3933	0.4834	0.644	0.595	-17.4633	-15.1923
ARIMA (2,1,0) (0,1,1) s	10.2902	0.4740	0.5022	0.613	0.575	-12.4988	-9.0923
ARIMA p = (1,3,4) d = (1,12) q = (1,2,4)	6.7751	0.2946	0.4043	0/745	0.671	-19.9113	-18.8984
ARIMA p = (1,4) d = (1,12) q = (1,4)	6.9792	0.3034	0.4157	0/738	0.696	-19.4289	-16.8869
ARIMA p = (4) d = (1,12) q = (4)	8.5717	0.3727	0.4252	0/678	0.663	-18.7016	-15.4306
ARIMA p = (1,3,4) d = (1,12) q = (4)	7.6240	0.3315	0.3868	0/714	0.668	-17.3962	-12.8542
ARIMA p = (1,3,4) d = (1,12) q = (1,3,4)	6.2836	0.1745	0.3377	0.837	0.761	-23.8404	-19.8382
ARIMA p = (4) d = (1,12) q = (1,3,4)	6.9822	0.3036	0.4064	0.739	0.696	-194189	-14.8769
ARIMA p = (1,4) d = (1,12) q = (1,3,4)	6.9721	0.3031	0.4116	0.738	0.680	-17.4523	-11.7748

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$$
$$= (1 - 0/059\, B - 0/27405\, B^3 - 0/48445\, B^4)$$
$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$$
$$= (1 - 057771\, B - 0/33765\, B^3 - 0/1263\, B^4)$$

Validation of the Model

Validation of the model already was done by statistical testing (Table 7), but for more accuracy, we developed the model with 30 of the factor data, and then predicted future value for next six months. The results of predictions and actual data are shown in Table 9. If we compare actual and prediction data, there are not significant differences between them, which may be prove the validation of the model. The results may show that Box-Jenkins time series may be applied to factor data instead of many water quality parameters and prediction of future variation of factors may help water quality management.

Table 8: Estimation of model's parameters

Model Parameter	*Estimate*
MA factor 1 lag 1	0. 57771
MA factor 1 lag 3	0.33765
MA factor 1 lag 4	-0.12630
AR factor 1 lag 1	0.05900
AR factor 1 lag 3	0.27405
AR factor 1 lag 4	-0.48445
Model Variance (sigma squared)	0.39214

Conclusion

The results of this work can be summarized as follows. Firstly, the results show dependency between water quality parameters: many of the correlation coefficients between parameters are above 0.3. This means that applying univariate statistical time series models for each parameter is not appropriate. Secondly, factor analysis with four-factor loading was identified as suitable for the interpretation of the water quality data. Thirdly, TC and TDS parameters are first factor loading. This means that contribution of ground water flow may be significant in increasing the TDS in the river. FC and TC rank as the second most important factor, because of discharge of domestic waste water from populations in the mountainous area of the basin that have no access to proper disposal method. Their high densities may arise from the use of fertilizer by farmers in the basin, and in the last factor BOD, COD and NH_3 are considerable. This means that the industrial and domestic wastewater also contribute in river water pollution. The result of developing Box-Jenkins for the factor data was reasonable and accurate, which means we may develop models for factor data instead of many water quality parameters. Finally, the result of this work may be helpful for water quality management of the river and similar places.

Table 9: Comparisons between actual and prediction data

Row	*Actual data*	*Prediction data*	*Lower confident interval*	*Upper confident interval*
31	16.9894	16.8125	15.7714	17.8536
32	15.2294	14.5900	14.5718	16.2541
33	157827	15.6845	14.6434	16.7257
34	16.2756	16.1026	14.9385	17.0208
35	15.9793	16.7966	15.7555	17.8378
36	15.3387	15.1318	14.0907	16.1730

References

Asadollahfardi, G. (2003). Analysis of surface water quality in Tehran. *Water Quality Research Journal of Canada*, **37(2):** 489-511.

Alberto, W.D., Del Pllar, D.M., Valeria, A.M., Fabiana, P.S., Cecillia, H.A. and B.M. Los Angles (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River Basin (Cordoba-Argentina). *Water Research J.*, **35(12):** 2881-2894.

Buckley, D.E. and G.A. Winter (1992). Geochemical characteristics of contaminated surflcial sediment in Halifax Harbour: Impact of waste discharge. *Canadian J. of Earth Science*, **29:** 2617-2639.

Box, G.E.P. and G.M. Jenkins (1976). Time series analysis, forecasting and control. Holden-Day, San Francisco, California, U.S.A.

Browne, M.W. (1968). A comparison of factor analytic techniques. *Pyschometrika*, **33:** 267-334.

Cary, N.C., Tabaching, B.G. and L.S. Fidell (2001). Using multivariate statistic, 4th edition. Allyn and Bacon, USA.

Hakstian, A.P., Rogers, W.T. and R.B. Cattle (1982). The behavior of numbers of factors with simulated data. *Multivariate Behaviour Research J*., **17(2):** 193-219.

Huck, P.M. and G.J. Farquhar (1974). Water quality models using Box and Jenkins method. *J. Environmental Engineering Division ASCE*, **100:** 733-753.

Iscen, C.F., Emiroglu, O., Iihan, S., Arslan, N., Yilmaz, V. and S. Ahiska **(**2008). Application of multivariate statistical techniques in the assessment of the surface water quality in Uluabat Lake Turkey. *Environmental Monitoring Assessment*, **144:** 269-276.

Johnson, R.A. and D.W. Wichern (2007). Applied multivariate statistical data analysis. 6th edition. Pearson Prentice Hall Publisher.

Jayawardena, A.W. and F. Lai (1987). Time series analysis of water quality data in Pearl River. *J. Environmental Engineering Division, ASCE*, **113:** 590-606.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Education Psychology Meas*., **20:** 141-151.

Kazi, T.G., Arian, M.B., Jamali, M.K., Jalbani, N., Afrd, H.I., Sarfraz, R.A., Baig, J.A. and Th. Kouimtzis (2003). Assessment of surface water quality in Northen Greece. *Water Research*, **37:** 4119-4124.

Lohani, B.N. and M.M. Wang (1987). Water quality data analysis in Chung Kang River. *J. Environmental Engineering Division ASCE*, **113:** 186-195.

Linn, R.L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, **33:** 37-71.

Mazlum, N., Ozen, A. and S. Mazlum (1999). Intepretation of water quality data by principal components analysis. *J. of Engineering and Environmental*, **23:** 19-26.

Ouyang, Y., Nkedi-Kizza, R., Wc, Q.T., Shinde, D. and C.H. Hung (2006). Assessment of seasonal variation in surface water quality. *Water Research J*., **40:** 3800-3810.

Perona, E., Bonilla, I. and P. Mateo (1999). Spatial and temporal change in water quality in a Spainish River. *The Science of the Total Environment*, **241:** 75-90.

Padro, R., Barrado, E., Cartilage, Y., Velasoco, M.A. and M. Vega (1993). Study of the contents and speciation of heavy metal in river sediments by factor analysis. *Anal. Lett*., **26:** 1719-1739.

Qiming, C., Xiyun, G., Yauwe, C. and M. Shengwei (1996). Dynamic variations of water quality in Tahia Lake and multivariate analysis of its influential factors. *Chinese Geographical Science J*., **6(4):** 364-374.

Renwick, W.H., Vanni, M.J., Zhang, Q. and J. Patton (2006). Water quality trends and changing agricultural practice in Midwest, U.S. watershed 1994-2006. *J. Environmental Quality*, **37(5):** 1862-1874.

Shah, A.Q. (2009). Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. *Ecotoxicology and Environmental Safety*, **72:** 301-309.

Singh, K.P., Malik, A. and S. Sinha (2005). Water quality assessment and apportionment of pollution source of Gomti River (India) using multivariate statistical techniques: A case study. *Analytica chimica Acta*, **538:** 355-374.

Shresthu, S. and F. Kazama (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River basin. *Japan Environmental Modelling and software*, **2:** 464-475.

Simeonv, V., Straitis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. and J. Schober (1985). Iron-oxo-hydroxides and their significance to the behaviour of heavy metals in estuaries. *Environmental Technology Lett*., **6(50):** 189-202.

SAS Institute Inc (1993). SAS/ETS Users Guide, version 6, 2nd ed.

Tucker, L.R., Koopman, R.F. and R.L. Linn (1969). Evaluation of factors analytic research procedures by means of simulated correlation matrix. *Psychomertrica*, **34:** 421-459.

Vega, M., Pardo, R., Barrado, E. and L. Deban (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research J*., **32(12):** 3581-3592.

Zeng, X. and T.C. Rasmussen (2005). Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA. *Journal of environmental quality*, **34(6):** 1980-1991.