Pol. J. Environ. Stud. Vol. 25, No. 3 (2016), 1149-1160
DOI: *10.15244/pjoes/61850*

Original Research

Application of Multivariate Statistical Methods for Water Quality Assessment of Karasu-Sarmisakli Creeks and Kizilirmak River in Kayseri, Turkey

Özgür Özdemir*

Malatya Metropolitan Municipality, Water and Sewerage Administration General Directorate (MASKI), Malatya, Turkey

Received: 29 January 2016
Accepted: 18 February 2016

Abstract

Monitoring water quality of rivers has particular ecological, economic, and environmental importance for a region. However, long-term monitoring of a river network requires a large amount of analysis with high work load and cost. In this study, multivariate analysis was performed to decrease the number of parameters and sampling points for monitoring water quality in the Kızılırmak River and its southern ranches of Karasu and Sarmısaklı creeks. On the study area, water quality parameters were collected from seven sampling points for 2004-07 and 2010-14. Multivariate statistical analysis was performed by hierarchical cluster analysis and principal component analysis for the evaluation of parameter variations and dataset interpretation. The dataset was divided into periods A and B to evaluate the effects of seasonal changes on sampling points and parameters.

Keywords: water quality, multivariate statistical analysis, PCA, HCA

Introduction

Water is an essential component of living organisms, and access to safe water is a main struggle for all life forms. Along with other freshwater sources, rivers are the main surface water sources for ecosystem life, agricultural use, industrial purposes, and drinking water production. On the other hand, rivers are vulnerable to pollution by excessive discharges of waste streams from anthropogenic, industrial, and agricultural activities from settlements around these rivers [1]. In addition to interactions of pollutants, climatic factors, hydrological characteristics, sediment, and metabolism in the water along with spatial and temporal variations determine water quality in a river [2]. Along with supplying adequate safe water, sustainable management of water quality in rivers with a reliable and representative quality monitoring program is one of the main tasks of governing officials. In river studies, field surveys and physicochemical measurements of an entire river network are traditionally carried out through monitoring programs. However, long-term field surveys generate large data sets, increasing processing load and operational costs [3]. Moreover, collection and classification of a large amount of data leads to difficulties in evaluation and representation of results along with

*e-mail: ozgurozd@hotmail.com

development of effective management strategies for water resources [4].

In recent decades, advanced statistical and calculation methods have been developed to classify large data clusters into meaningful range, extract useful information, identify relationships among corresponding data, and evaluate the results. Although many methods such as projection pursuit technique and neural networks have been developed, multivariate statistical techniques have enabled better and easier assessment for interpretation of complicated data sets in water quality studies [5]. Similarly, another study stated that multivariate statistical techniques are useful for reducing large data obtained from physicochemical water quality studies [6]. Those techniques are also effective for river water classification, determination of temporal and spatial variations caused by natural and anthropogenic factors, and development of appropriate strategies for effective management of water resources [7-9]. Furthermore, multivariate statistical techniques have been applied to a variety of environmental issues, including risk assessment in wastewater management, evaluating water recycling strategies, and assessment of groundwater hydrology and chemistry [10-12]. Among multivariate statistical techniques, hierarchical cluster analysis (HCA), principal component analysis (PCA), discriminant analysis (DA), and principal factor analysis (PFA) are mostly applied methods for water quality investigations. In cluster analysis the data set is grouped on the basis of similarities and differences while dimensionality of a data set is reduced into a new set of variables with principal component analysis. Reduction generates new inter-related variables, and the principal components are arranged in descending order of importance for explaining variance of all original property.

Many researchers have applied HCA and PCA analyses for quality classification and monitoring of temporal variations in water systems [1]. For instance, the factors controlling arsenic mobilization in groundwater chemistry were investigated and groundwater areas were classified with these methods [13]. In another study, HCA and PCA were applied using 10 chemical variables in 247 samples to classify groundwater [14]. River water quality was analysed in the Three Gorges area of China to investigate spatial and temporal variations along with identification of potential pollution sources [15]. After multivariate analyses, researchers determined major pollution factors and obtained fundamental information to develop better water pollution control strategies. Multivariate statistical techniques of PCA and PFA were employed to determine important water quality parameters in the St. Johns River in the U.S.A. [16]. They reported that total organic carbon, dissolved organic carbon, total nitrogen, dissolved nitrate and nitrite, orthophosphate, alkalinity, salinity, Mg, and Ca were the most important parameters in assessing variations of water quality in the river. HCA, DA, and PFA were applied for the assessment of seasonal variations in the surface water quality of tropical pastures in Kuala Lumpur [17]. The study was conducted with a dataset consisting of one-year monitoring of 14 parameters at six sampling sites. The authors stated that multivariate techniques enabled them to plan for future sampling events, optimize the number of sampling points, select appropriate water quality parameters, and reduce the costs. HCA was applied to 12 months of data from eight sampling points based on seasonal differences and different levels of pollution [18]. Additionally, the application of PCA/PFA helped to evaluate seasonal and spatial variations in river water quality and identify corresponding pollution sources in the Kaduna River.

The objective of this study was to determine the variables responsible for spatial and temporal variations in water quality of Karasu and Sarmısaklı creeks, together with the Kızılırmak River, by applying multivariate statistical techniques. Two multivariate statistical methods of Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) were applied using STATISTICA 13 and XLStat software packages in order to interpret the relationships between parameters and sampling sites. All studies were performed using the datasets obtained during the observation period of 2004-07 and 2010-14.

Materials and Methods

Study Area and Sampling

The Kızılırmak flows for a total of 1,355 km, which rises in Eastern Anatolia around 39.8°N 38.3°E, passing to the northeast of Lake Tzu, and finally flows through Gokırmak Delta into the Black Sea at 41.72°N 35.95°E (Fig. 1) [19]. The water of Sarmısaklı Creek is mostly used for agricultural irrigation in the Kayseri plain. Sarmısaklı Dam is located almost 40 km east of the confluence point of Sarmısaklı and Karasu streams. Karasu Creek is highly polluted by discharging domestic and industrial streams from the city of Kayseri.

The water quality of rivers was monitored by periodically sampling seven points in 2004-08 and 2010-14. The locations of sampling points were as described in Table 1. In total, 196 samples were collected and analysed for BOD_5, COD, ammonia-N, kjeldahl-N, oil-grease, fecal and total coliform, TSS, TDS, nitrite, nitrate, total and ortho phosphate, surfactants, chloride,

Fig. 1. Location of study area in Turkey.

Table 1. Descriptions and coordinates of sampling locations.

Sampling Points	Description	SP Coordinates
SP1	on Sarımsaklı creek 300 m before confluence to Karasu creek	38°44'59.25"N, 35°18'51.65"E
SP2	on Karasu creek 400 m before confluence to Sarımsaklı creek	38°44'49.77"N, 35°18'39.26"E
SP3	on Karasu creek 100 m before wastewater treatment plant discharge point	38°45'52.37"N, 35°18'11.86"E
SP4	on Karasu creek 400 m after wastewater treatment plant discharge point	38°46'06.88"N, 35°17'42.91"E
SP5	on Karasu creek 100 m before confluence to Kızılırmak River	38°49'50.70"N, 35°13'14.87"E
SP6	on Kızılırmak river 300 m before Karasu creek confluence point	38°49'59.08"N, 35°13'22.52"E
SP7	on Kızılırmak river 300 m after Karasu creek confluence point	38°49'56.99"N, 35°12'53.23"E

Table 2. Average temperature and rainfall for Kayseri province and surroundings.

Parameter	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Average temp.(ºC)	-1.7	0.0	5.0	10.8	15.1	19.2	22.7	22.2	17.3	11.6	5.1	0.4
Average rainfall (kg/m^2)	33.5	35.9	42.0	52.3	52.3	40.1	10.1	5.9	13.9	28.5	33.5	39.1

Table 3. Water quality criteria for inland surface water resources.

Parameters	Class I	Class II	Class III	Class IV
Temp. (ºC)	≤ 25	≤ 25	≤ 30	> 30
pH	6.5-8.5	6.5-8.5	6.0-9.0	except 6.0-9.0
Conductivity (µS/cm)	< 400	400-1,000	1,001-3000	> 3,000
Dissolved O_2 (mg/L)	> 8	6-8	3-6	< 3
Saturated O_2 (%)	90	70-90	40-70	< 40
COD (mg/L)	< 25	25-50	50-70	> 70
BOD_5 (mg/L)	< 4	4-8	8-20	> 20
NH_4^+-N (mg/L)	< 0.2[b]	0.2-1[b]	1-2[b]	> 2
NO_2^--N (mg/L)	< 0.002	0.002-0.01	0.01-0.05	> 0.05
NO_3^--N (mg/L)	< 5	5-10	10-20	> 20
T. Kjeldahl-N (mg/L)	0.5	1.5	5	> 5
Total-P (mg/L)	< 0.03	0.03-0.16	0.16-0.65	> 0.65
Hg^{+2} (µg/L)	< 0.1	0.1-0.5	0.5-2	> 2
Cd^{+3} (µg/L)	≤ 2	2-5	5-7	> 7
Pb^{+2} (µg/L)	≤10	10-20	20-50	> 50
Cu^{+2} (µg/L)	≤20	20-50	50-200	> 200
Ni^{+2} (µg/L)	≤20	20-50	50-200	> 200
Zn^{+2} (µg/L)	≤200	200-500	500-2,000	> 2,000
Faecal Coliform (MPN/100 mL)	≤10	10-200	200-2,000	> 2,000
Total Coliform (MPN/100 mL)	≤100	100-20,000	20,000-100,000	> 100,000

sulfate, fluoride, boron, dissolved oxygen, cadmium, chromium, copper, mercury, nickel, lead, and zinc. All laboratory analyses were performed by following the instructions in the Standard Methods. On the other hand, temperature and pH were measured *in situ* by a multi-parameter.

Table 4. Summarized descriptive statistics for the raw data set.

Variables	Mean	Conf. 95%	Conf. -95%	Median	Range	Std. Dev.	Coeff. Var.	Std. Err.	Skewness
BOD_5 (mg/L)	29.46	21.23	37.69	10.00	357.00	57.99	196.82	4.17	3.89
COD (mg/L)	73.17	55.04	91.29	35.00	904.00	127.65	174.47	9.19	4.01
NH_4^+-N (mg/L)	3.83	2.03	5.64	0.85	36.28	7.95	207.46	0.91	2.83
T.Kjeldahl-N (mg/L)	5.81	4.50	7.13	2.38	76.76	9.14	157.31	0.67	4.10
Oil-Grease (mg/L)	8.17	6.94	9.39	10.00	52.97	8.64	105.80	0.62	2.90
F.Coliform (MPN/100 mL)	322.59	209.85	435.33	24.00	1,500.00	454.99	141.04	56.43	1.11
T.Coliform (MPN/100 mL)	1,076.45	96.10	2,056.79	40.00	16,000.00	2,853.88	265.12	482.39	4.60
TSS (mg/L)	43.12	34.24	52.01	26.00	461.00	62.58	145.11	4.50	4.20
TDS (mg/L)	1,151.14	1,072.72	1,229.56	1,125.00	2,164.00	345.51	30.01	39.37	0.08
NO_2^--N (mg/L)	0.85	0.50	1.20	0.15	15.99	2.27	267.39	0.18	4.07
NO_3^--N (mg/L)	1.70	1.43	1.96	1.10	8.59	1.77	104.44	0.13	1.55
Total-P (mg/L)	1.17	0.93	1.42	0.60	10.80	1.66	141.16	0.12	3.18
Ortho-P (mg/L)	1.45	1.07	1.84	0.62	16.12	2.44	167.45	0.20	3.93
Surfactants (mg/L)	0.57	0.45	0.69	0.42	7.01	0.83	146.52	0.06	4.83
Cl^- (mg/L)	298.44	240.15	356.74	238.00	846.96	256.84	86.06	29.27	0.43
SO_4^{-2} (mg/L)	174.94	157.61	192.26	143.56	949.91	122.00	69.74	8.78	2.14
F^- (mg/L)	0.63	0.48	0.77	0.32	6.73	1.00	158.78	0.07	3.73
B (µg/L)	18.52	3.79	33.25	1.10	848.00	104.00	561.55	7.47	6.92
DO (mg/L)	4.81	4.44	5.19	4.81	17.46	2.64	54.75	0.19	0.61
pH	7.46	7.39	7.52	7.40	3.10	0.45	5.99	0.03	0.74
Temp. (ºC)	18.17	17.51	18.82	18.40	24.10	4.63	25.47	0.33	-0.64
Cd^{+3} (µg/L)	2.70	0.51	4.89	0.02	171.00	15.44	572.11	1.11	8.50
Cr^{+3} (µg/L)	7.21	0.39	14.03	0.03	650.00	48.16	668.17	3.46	12.56
Cu^{+2} (µg/L)	3.89	2.73	5.06	0.02	45.60	8.24	211.70	0.59	1.96
Hg^{+2} (µg/L)	1.57	0.91	2.23	1.00	32.72	4.50	286.60	0.33	5.40
Ni^{+2} (µg/L)	21.45	-7.24	50.14	0.07	2,790.00	202.61	944.57	14.55	13.40
Pb^{+2} (µg/L)	1.71	1.10	2.31	0.08	29.99	4.29	251.75	0.31	3.67
Zn^{+2} (µg/L)	15.21	8.88	21.55	0.16	308.00	44.00	289.20	3.21	4.47

Average precipitation and temperature values for the Kayseri region were as given in Table 2. Precipitation between December-May was higher than the period between June-November, thus statistical evaluations were performed for two main periods: Period A, for data obtained in winter and spring seasons (December-May) and Period B, for data gathered during summer and autumn (June-November).

Water quality of rivers was evaluated according to Turkish surface water quality management regulations [20]. In the regulation, surface waters are divided into four classes as shown in Table 3. Class I is clean water that can be used for recreational purposes and irrigation and domestic purposes after disinfection. Class II is fairly clean water that can be used for recreational purposes, fish farming, and drinking after treatment. Class III is polluted water that can only be used as for industrial purposes after treatment, while Class IV is heavily polluted water that should not be used at all. In our study, threshold values of Class II and Class III were chosen as reference for parameter assessment.

Table 5. Ranges of selected parameters in seven sampling points over 2004-07 and 2010-14.

Parameters	SP1	SP2	SP3	SP4	SP5	SP6	SP7
BOD_5 (mg/L)	5-360	5-195	5-85	5-52	5-60	3-20	5-46
COD (mg/L)	10-914	10-257	10-148	10-132	12-124	10-338	10-148
NH_4^+-N (mg/L)	0.69-36.3	0.17-6.5	0.2-5.01	0.06-3.3	0.05-2.14	0.02-2.87	0.04-2.85
T.Kjeldahl-N (mg/L)	1-76.8	0.04-17	0.83-19.5	1.1-19.2	1-17	0.14-12	0.14-21
Oil-Grease (mg/L)	0.07-52.5	0.07-10.5	0.03-53	0.03-29.5	0.03-16	0.03-35.5	0.03-26
F.Coliform (MPN/100 mL)	11-1,250	0-1,200	0-1,100	0-1,100	2.3-1,500	0-1,100	0-1,100
T.Coliform (MPN/100 mL)	11-1,100	15-5,400	4.3-1,100	2.4-1,100	4.3-16,000	0-1,100	2.3-1,100
TSS (mg/L)	9-463	5-106	10-96	5-130	6-70	2-52	7-122
TDS (mg/L)	508-2,318	174.1-1,784	978-1,760	916-1,520	1,126-1,561	154-1,401	809-1,310
NO_2^--N (mg/L)	0.01-16	0.01-10	0.03-10	0.02-10	0.09-4.74	0.01-10	0.01-10
NO_3^--N (mg/L)	0.01-8.05	0.21-5.93	0.11-5.9	0.05-5.5	0.01-6.8	0.01-7.3	0.01-8.6
Total-P (mg/L)	0.08-10.82	0.12-4.94	0.06-2.5	0.15-5.6	0.05-5.5	0.02-2.1	0.13-2.11
Ortho-P (mg/L)	0-16.12	0.08-6.45	0.05-1.79	0.08-5.17	0.11-3.02	0.02-1.5	0.04-2.62
Surfactants (mg/L)	0-7.02	0-2.12	0.01-1.83	0-1.23	0.01-1.46	0-1.5	0-1.91
Cl^- (mg/L)	0.27-777	0.09-802	0.04-847	0.09-749	401-600	0.05-660	0.07-445
SO_4^{-2} (mg/L)	9.09-367	26.13-433	20.58-515	27.55-310	28.77-321	0.09-650	0.28-950
F^- (mg/L)	0-6.73	0.01-2.3	0-6	0-6	0-3.21	0-2.54	0-1.7
B^+ (mg/L)	0-274	0.05-204	0.03-848	0-823	0-716	0-105	0-341
DO (mg/L)	0.01-7.2	0.01-8.6	0.01-10.67	0.02-9.33	0.02-7.13	0.34-17.47	0.16-12.58
pH	6.4-9.5	6.71-8.18	6.56-7.91	6.51-8.79	6.8-7.7	6.99-8.3	6.81-8.58
Temp. (ºC)	8.2-28.6	9-26.3	5.3-26.6	7.9-24.3	9.2-24.3	4.5-25	7.28-25.3
Cd^{+2} (µg/L)	0-171	0-2.25	0-32	0-57	0-62	0-2.09	0-23
Cr^{+3} (µg/L)	0.01-650	0.01-136	0.01-43.3	0-31.6	0-42.17	0-25.7	0-22.8
Cu^{+2} (µg/L)	0.01-45.6	0.01-20	0-20	0.01-20	0.01-20	0.01-20	0.01-20
Hg^{+2} (µg/L)	0-4.81	0-11.42	0-22.31	0-21	0-2.95	0-32.72	0-6.07
Ni^{+2} (µg/L)	0.01-2790	0.03-442	0.01-60.2	0-45.6	0-55.5	0-30.8	0-28
Pb^{+2} (µg/L)	0.01-30	0.01-15.63	0.01-14.3	0.01-14.98	0.01-17.5	0.01-8	0.01-13.97
Zn^{+2} (µg/L)	0-308	0-113	0-58	0-106	0-271	0-95	0-200

Data Preparation

In multivariate analysis variables should be standardized since variables with different scales and units may not contribute equally to the statistical analysis. Transforming the data to comparable scales can prevent this problem by minimizing the influence of variable variance, eliminate the influence of different units of measurement, and render the data dimensionless [21]. This can be achieved by calculating the standard scores (z-scores) of the raw data set by using the following equation [22].

$$z_{ij} = \frac{x_{ij} + x_j}{s_j} \quad (1)$$

Some researchers also recommend a log-transformation of raw data set before standardization in order to maximize the effectiveness of statistical analysis [23]. In our study, data screening showed that all data except temperature was skewed positively for all parameters (Table 4). Therefore, statistical analyses were carried out after standardizing the logarithmic transformed data set.

Table 6. Mean values of parameters in sampling points.

Parameters	SP1 Class III	SP2 Class III	SP3 Class II	SP4 Class III	SP5 Class III	SP6 Class III	SP7 Class III
BOD_5 (mg/L)	102.45	20.41	21.57	16.76	21.27	8.48	12.41
COD (mg/L)	237.24	50.96	45.32	47.86	55.32	31.52	37.14
NH_4^+-N (mg/L)	18.95	1.19	1.38	1.21	0.79	0.56	0.99
Kjeldahl-N (mg/L)	15.48	4.04	4.46	5.16	4.64	2.91	3.64
Oil-Grease (mg/L)	13.81	6.18	8.48	6.91	7.62	7.26	6.63
F.Coliform (MPN/100 mL)	589.5	246.03	142.03	194.79	679.72	172.4	412.24
T.Coliform (MPN/100 mL)	533	1369	319.66	412.78	4,345.66	328.8	226.22
TSS (mg/L)	111.55	33.3	40.54	35.93	35.73	12.97	29.31
TDS (mg/L)	1,026.92	1,012.59	1,416.67	1,258.67	1,390	979.08	1,113.42
NO_2^--N (mg/L)	1.1	0.98	0.85	0.89	0.54	0.73	0.8
NO_3^--N (mg/L)	1.49	2.12	1.55	1.66	1.73	1.57	1.77
Total-P (mg/L)	2.42	1.54	0.58	1.4	1.26	0.31	0.74
Ortho-P (mg/L)	3.63	1.82	0.61	1.67	1.03	0.38	0.92
Surfactants (mg/L)	1.26	0.41	0.45	0.42	0.62	0.36	0.45
Cl^- (mg/L)	140.66	354.75	441.29	354.4	505.03	205.26	208.23
SO_4^{-2} (mg/L)	130.3	138	136.19	137.47	146.11	289.29	236.35
F^- (mg/L)	0.91	0.54	0.63	0.73	0.51	0.57	0.46
B^+ (mg/L)	9.94	10.13	32.05	30.22	34.14	3.8	12.55
DO (mg/L)	3.28	4.86	4.73	4.61	3.78	6.59	5.59
pH	7.72	7.3	7.21	7.3	7.25	7.8	7.55
Temp (ºC)	19.41	18.84	18.13	18.58	19.24	16.31	16.98
Cd^{+2} (µg/L)	9.9	0.23	1.65	2.36	3.37	0.16	1.22
Cr^{+3} (µg/L)	27.96	8.72	2.43	2.73	4.22	1.48	2.3
Cu^{+2} (µg/L)	5.12	4.45	2.86	3.49	4.58	3.49	3.49
Hg^{+2} (µg/L)	0.85	1.41	2.16	1.59	0.96	2.91	0.93
Ni^{+2} (µg/L)	102.61	20.88	3.48	5.91	8.75	1.83	3.58
Pb^{+2} (µg/L)	3.75	1.79	1.24	1.19	1.81	0.73	1.46
Zn^{+2} (µg/L)	33.88	11.33	7.26	13.44	23.3	6	12.75

Statistical Analyses

Hierarchical cluster analysis (HCA) is a combination of techniques to classify large data into clusters on the basis of similarities or dissimilarities. Thus, resulting groups are similar to each other but distinct from other groups. Researchers have widely applied HCA for classification of water quality and interpreting experimental data [24-25]. In this study, HCA was used to group sampling points based on their similarities in water quality and to detect links between water quality parameters. A combination of linkage methods of Ward, single, complete and distance methods of Euclidian and squared Euclidian were applied to all data sets. PCA analysis is widely used to assess spatial and temporal variations in water quality [7, 26]. In this study, PCA was performed in order to evaluate the significance of parameters in water quality assessment and to determine the correlation between parameters. Before PCA analysis, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's Test of Sphericity (p-value) were applied on experimental data to test partial correlation and dependency for excluding the potential non-independent data that can affect PCA. The Kaiser criterion was applied to determine the total number of factors for each dataset. Factors with eigenvalues greater than or equal to 1 were accepted as possible sources of variance in the data, with the highest priority ascribed to the factor that has the highest eigenvector sum.

The reason for choosing 1 was that a factor must have a variance at least as large as that of a single standardized original variable to be acceptable [27]. All multivariate analyses were performed using STATISTICA 13 and XLStat software packages.

Results and Discussion

Characterization of Surface Water Quality

Water quality of Sarımsaklı and Karasu creeks of the Kızılırmak River around Kayseri Province were monitored at seven sampling points over a period of 10 years. Variations in chemical, physical and microbiological parameters are as shown in Table 5, while mean values are given in Table 6. The values in the tables indicate that SP1 on Sarımsaklı Creek has significantly higher values of organic pollutants and most of other parameters compared to other sampling locations. These high-pollutant concentrations indicated that SP1 had lower water quality than other sampling locations. On the other hand, other locations had a homogenized fluctuation for the parameters.

Water classification of working areas were performed according to values of COD, BOD_5, ammonium-N, nitrate, total-P, chloride, sulphate, and TDS, and the results are summarized in Table 7. In this table, Sarmısaklı and Karasu creeks were evaluated as a combined system according to quality classes. All parameter values except for SP6 and SP7 imply that water quality was Class III or Class IV. Water quality class in some parts of Sarmısaklı Creek was changing from Class II to Class IV due to the discharge of domestic and industrial wastewater from the Kayseri region. Some parameter values (including especially BOD_5, COD, Kjeldahl-N, nitrite, total-P, DO, pH, and Hg) were exceeding the water quality criteria of Class II and Class III. Higher nitrogen and phosprous loading at Sarmısaklı Creek was associated with the discharge from agricultural activities, cattle breeding, and other land activities. On the other hand, water quality of Sarmısaklı Creek according to nitrate (Class I or Class II) indicates that the pollution did not originate from agricultural sources.

Hierarchical Cluster Analysis

HCA was performed at group sampling sites based on the similarities in water quality parameters. Among the methods applied for linkage and distance measurement, the Euclidian distance with Ward's method provided the best results. Based on 28 variables, the HCA produced a dendogram grouping all seven sampling points into three statistically significant main clusters as shown in Fig. 2. Cluster 1 contains SP6 and SP7, which were located on the Kızılırmak River. Cluster 2 consists of SP2, SP3, SP4, and SP5, which have approximately the same water quality. SP3 and SP4 have more similar features compared to SP5 and SP2. Cluster 3 contains only SP1, which has the worst water quality among all sampling points. Cluster 1 and Cluster 2 sampling points have more similar features compared to Cluster 3. On the other hand, SP2 is improving the water quality after confluence with SP1.

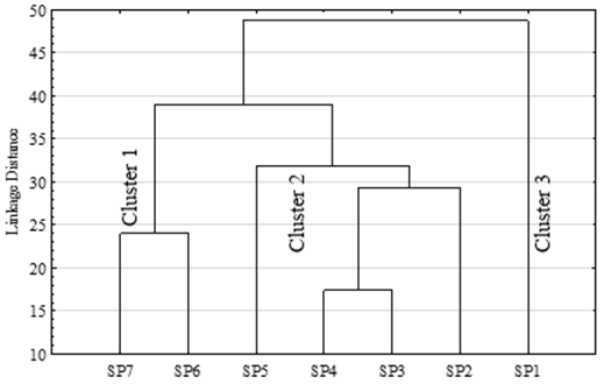

Fig. 2. HCA results of sampling sites according to water quality parameters for all periods.

Datasets for Period A and Period B were analysed separately with HCA in order to examine characteristics of sampling points during seasonal changes in water quality. The result of analysis is as shown in Fig. 3. In contrast to Fig. 2, Cluster 1 consisted only of SP6 and SP7 joined to Cluster 2 in Period A. Cluster 3 contains only SP3 in all periods with its specific water quality. These results indicate that water quality of sampling points during Period A was affected by the variations in seasonal precipitation. Those affections were mainly associated with the flow of additional rainfall during that period. On the other hand, no significant changes were obtained with the HCA analysis in Period B. HCA analysis implied that there were three separate water qualities in the studied area. Overall evaluation revealed that the HCA method is useful for classification of water quality and it showed that several sampling points can be decreased with this method and only one station from each cluster could be used for rapid spatial assessment of water quality in whole network.

Table 7. Water quality classes according to several parameters.

Location	COD (mg/L)	BOD_5 (mg/L)	NH_4^+-N (mg/L)	NO_3^--N (mg/L)	Total-P (mg/L)	Cl^- (mg/L)	SO_4^{-2} (mg/L)	TDS
Kızılırmak River	I-II	II	II-IV	I-II	II-IV	III-IV	III-IV	II-III
Sarımsaklı-Karasu Creek	II-IV	II-IV	III-IV	I	III-IV	II-IV	I-IV	II-III

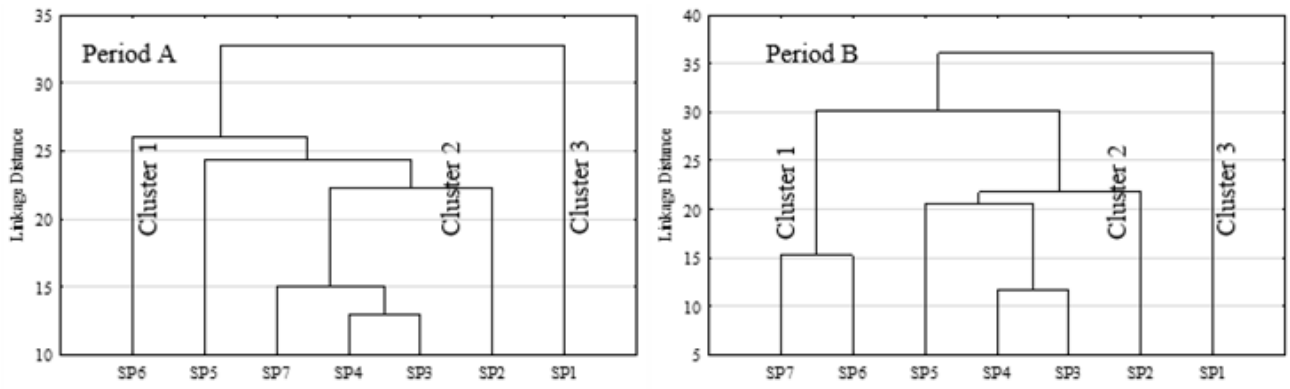

Fig. 3. HCA results of sampling sites according to water quality parameters for Period A and Period B.

HCA was applied to the entire water quality data set and the results are as shown in Fig. 4. Among the applied methods for linkage and distance combination, the best results for variable relation were obtained using Ward's method and Euclidian distance. According to analysis, parameters were roughly classified into two major groups: Cluster 1 ranged from Cl to ammonia-N and Cluster 2 from Zn to BOD_5. However, clusters can be further clustered in five groups in detail analysis. Cluster 1 contained Cl^-, temperature, DO, SO_4^{-2}, and TDS as variables while Cluster 2 had the parameters of surfactants, Hg, and oil-grease. Cluster 3 consisted of F, NO_3^--N, ortho-P, Total-P, F.Coliform, TSS, T. Kjeldahl-N, and NH_4^+-N. Cluster 4 included the heavy metal pollutants of Zn^{+2}, Cu^{+2}, Ni^{+2}, Cr^{+2}, Pb^{+2}, and Cd^{+2}. Finally, Cluster 5 consisted mostly of biological-organic parameters such as those of BOD_5, COD, T. Coliform, B, NO_2^--N, and pH variables. The variables in each cluster were linked at different levels of similarity. The closest similarity was observed between Ni^{+2} and Cr^{+3} within Cluster 4, whereas the second closest linkage was between COD and BOD_5 in Cluster 5.

HCA results for quality parameters for Period A and Period B were as shown in Fig. 5. For Period A, all parameters were classified into two main clusters: Cluster 1 from F to ammonium-N and Cluster 2 from Zn to BOD_5. Two main clusters were further grouped into five clusters. Cluster 1 contained F^-, Cl^-, DO, NO_3^--N, SO_4^{-2}, and TDS, while Cluster 2 included temperature, Hg and oil-grease. Cluster 3 consisted of NO_3^--N, surfactants, B, TSS, Ortho-P, total-P, T. Coliform, F. Coliform, T. Kjeldahl-N, and NH_4^+-N. Cluster 4 includes Zn^{+2}, Cu^{+2}, Ni^{+2}, Cr^{+3}, Pb^{+2}, and Cd^{+2}. Cluster 5 consisted of BOD_5, COD, and pH variables. The general pattern for Period A was consistent with results obtained for the whole period with only some minor changes. The closest similarity was between Ni^{+2} and Cr^{+3} within Cluster 4 similar to the results of HCA of the whole dataset. COD/BOD_5, Pb^{+2}/Cd^{+2}, and $T.Kjeldahl/NH_4^+$-N were other parameter couples with close similarities in different clusters.

For Period B, all parameters could be classified into two main groups: Cluster 1 had parameters from Cl^- to F. Coliform while Cluster 2 had Zn^{+2} to BOD_5. Five clusters were created from two major clusters. Cluster 1 contained Cl^-, temperature, DO, SO_4^{-2}, F. Coliform, and TDS, and Cluster 2 consisted of surfactants, Hg^{+2}, and oil-grease. Cluster 3 had F^-, NO_3^--N, ortho-P, total-P, TSS, Kjeldahl-N, and NH_4^+-N, and Cluster 4 included Zn^{+2}, Cu^{+2}, Ni^{+2}, Cr^{+6}, Pb^{+2}, and Cd^{+2}. Finally, Cluster 5 consisted of B, NO_2^--N, T. Coliform, BOD_5, COD, and pH parameters. As distinct from Period A, Cluster 4 was not related closely to Cluster 5.

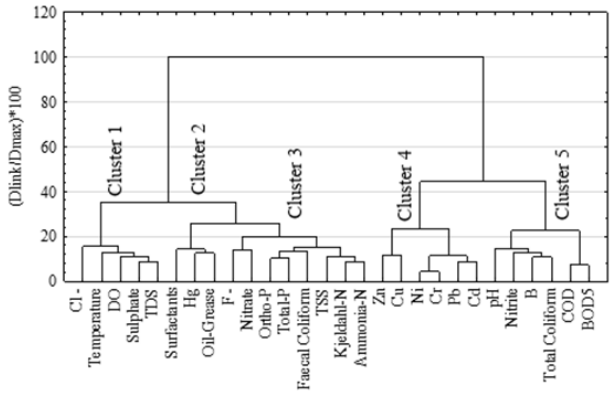

Fig. 4. HCA clustering of parameters for all periods.

Principal Component Analysis

In order to perform the PCA, the software was initially run unrotated for 28 parameters. PCA with varimax rotation was also applied but minor changes appear with this rotation. Therefore, varimax rotation results are not discussed. PCA enables us to compare the parameter relations by the automatically produced correlation matrix and observe the relationships between the parameters. The correlation matrix obtained from first run of PCA showed that some parameters were not significantly correlated with other parameters in the dataset. Therefore, parameters of fecal coliform, total coliform, ortho-P, TDS, NO_2^--N, NO_3^--N, Cl^-, SO_4^{-2}, F^-, B^+, pH, and temperature were removed from both Period A and Period B datasets for further PCA analysis. Additional PCA applications were performed by using the remaining 16 water-quality parameters.

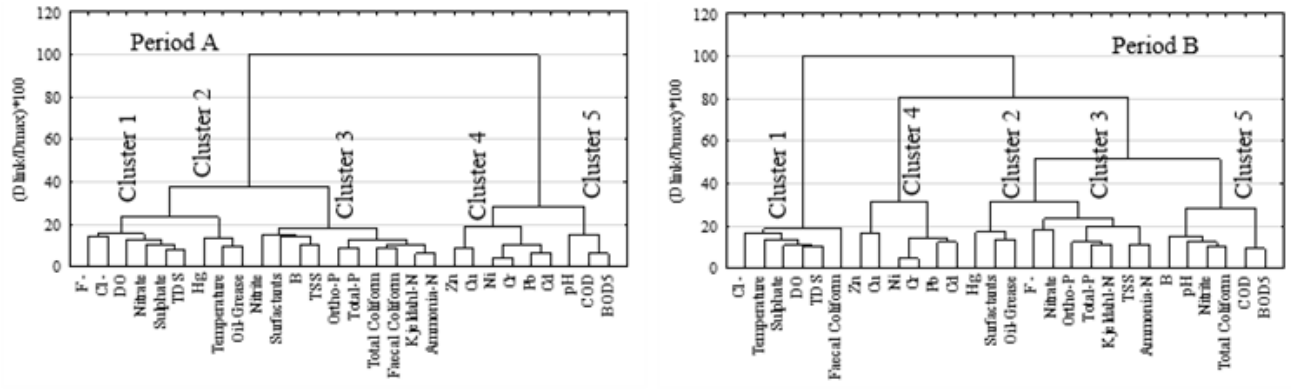

Fig. 5. HCA clustering of parameters for Period A and Period B.

Additionally, the matrix shows that BOD_5, COD, NH_4^+-N, and T. Kjeldahl-N parameters have high correlations with each other for Period A and Period B datasets.

Kaiser-Meyer-Olkin (KMO) values were calculated as 0.792 and 0.724 for Period A and Period B, respectively. These KMO values indicate that PCA could achieve a significant reduction in the dimensionality of the original dataset. Bartlett's sphericity test for both periods was significant ($p< 0.05$). The results of eigenvalues for both periods are shown in Fig. 6. According to these results, the first four PCs shown in the scree plot are extracted for further evaluation in Period A, while three PCs were extracted from the factor loadings in Period B.

The relationship between extracted PCs and parameters along with eigenvalues and PC variances for Period A and Period B are given in Table 8. Factor loadings are generally classified as strong, moderate, and weak, corresponding to absolute loading values of >0.75, 0.75-0.50, and 0.50-0.30, respectively [26]. In this study, values greater than 0.5 are accepted as effective loadings. The extracted 4 PCs for Period A were representing a total variance of 78.8%. Based on the PC loadings, 16 parameters were grouped into four PC groups. PC1 contains Surfactants, Cd^{+2}, Cr^{+6}, Cu^{+2}, Ni^{+2}, Pb^{+2}, Zn^{+2}, and PC2 of BOD_5, COD, NH_4^+-N, T. Kjeldahl-N, TSS, and Total-P. PC3 includes oil-grease and Hg^{+2}, while PC4 had only DO.

PC1 accounts for a total variance of 34.2%, and all parameters except surfactants have a negative loading on this component. PC2 accounts for a total variance of 25.9% and has a negative loading with BOD_5, COD, NH_4^+-N, Kjeldahl-N, TSS, and Total-P. PC3 with two parameters represents a total variance of 11.6% with high negative effect. The last component of PC4 accounts for 6.9% of total variance in Period A and contains a positive loading only with DO.

Three PCs were extracted in Period B, with 68.8% variance of raw dataset (Table 8). Sixteen parameters for Period B were divided into three PC groups. The PC1 included with significant high factor loadings of the parameters BOD_5, COD, NH_4^+-N, T. Kjeldahl-N, oil-grease, TSS, Total-P, surfactants, Cd^{+2}, Cr^{+6}, Cu^{+2}, Ni^{+2}, and Pb^{+2} PC1 accounted for 34.7% of total variance, and the parameters of Cd^{+2}, Cr^{+6}, Cu^{+2}, Ni^{+2}, and Pb^{+2} had positive loading on this PC. The second PC, with 23.6% of total variance, contained the parameters of DO, Cd^{+2}, Cr^{+6}, Ni^{+2}, and Pb^{+2} with high factor loadings. Only DO has a negative loading on this PC. The last PC for Period B accounted for a total variance of 10.6% and contained the parameters of oil-grease and Hg^{+2} with high negative loadings.

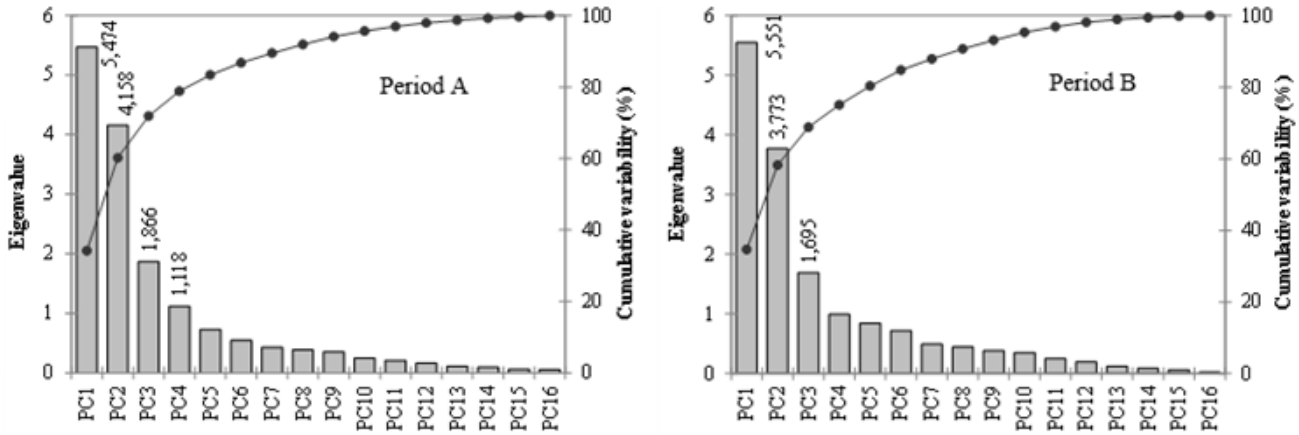

Fig. 6. Screen plot of eigenvalues versus water quality data for Period A and Period B.

Table 8. Eigenvalues and PC (factor) loadings of PCA for Period A and Period B dataset (Eigenvalue>1).

Parameters	Period A				Period B		
	PC1	PC2	PC3	PC4	PC1	PC2	PC3
NH_4^+-N (mg/L)	0.243	-0.816	0.049	0.242	-0.535	0.447	0.325
BOD_5 (mg/L)	-0.045	-0.887	-0.070	0.120	-0.707	0.483	0.146
Cd^{+2} (µg/L	-0.927	-0.120	0.173	-0.086	0.676	0.646	0.090
COD (mg/L)	-0.148	-0.889	0.070	-0.021	-0.668	0.456	0.270
Cr^{+3} (µg/L	-0.926	-0.081	-0.225	0.023	0.602	0.719	-0.153
Cu^{+2} (µg/L	-0.853	0.017	-0.148	-0.013	0.634	0.302	0.176
DO (mg/L)	-0.080	0.429	0.024	0.789	0.374	-0.679	0.064
Hg^{+2} (µg/L	0.143	-0.026	-0.917	-0.050	-0.196	0.224	-0.761
Kjeldahl-N (mg/L)	-0.070	-0.631	0.350	0.451	-0.636	0.304	0.403
Ni^{+2} (µg/L	-0.905	-0.169	-0.117	-0.076	0.615	0.672	-0.130
Oil-Grease (mg/L)	0.284	-0.375	-0.734	0.306	-0.590	0.129	-0.671
Pb^{+2} (µg/L	-0.905	-0.056	0.035	0.072	0.601	0.683	-0.065
Surfactants (mg/L)	0.641	-0.490	-0.250	-0.216	-0.630	0.281	-0.442
Total-P (mg/L)	0.148	-0.568	0.319	-0.158	-0.611	0.437	0.091
TSS (mg/L)	-0.144	-0.757	0.021	-0.174	-0.650	0.315	0.101
Zn^{+2} (µg/L	-0.862	-0.047	-0.265	-0.026	0.485	0.447	0.052
Eigenvalue	5.474	4.158	1.866	1.118	5.551	3.773	1.695
Variability (%)	34.21	25.988	11.664	6.985	34.694	23.584	10.592
Cumulative %	34.21	60.198	71.862	78.847	34.694	58.278	68.87

Another option to observe the correlations of parameters with obtained factor loadings is the graphical presentation. The relationship of parameters with main PC loadings (Factor1 and Factor2) for Period A and Period B are represented in Fig. 7. For both periods, parameters of DO and Hg were always outliers. The heavy metals in Period A are slightly better correlated than in Period B.

The results of PCAs for both periods are compared with the results of previous HCA analysis outputs. Comparison indicated that the results of PCA and HCA were not consistent with each other. Therefore, 16 significant parameters from PCA were grouped by HCA. Several linkage methods and distance configurations were applied for the clustering analysis. The best results were observed with single linkage and Euclidian distance combinations. The results of both periods of HCA were as shown in Fig. 8. It can be clearly seen that obtained groupings are fully consistent by comparing relevant periods.

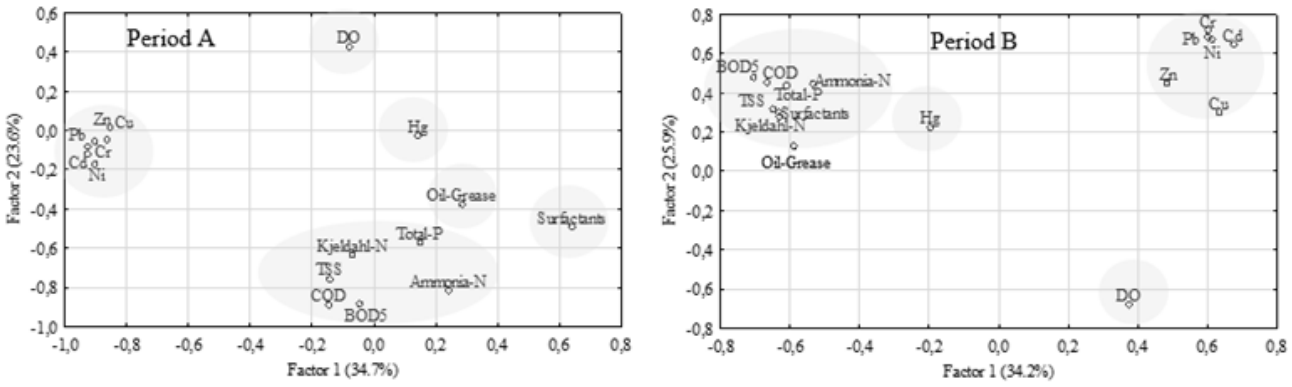

Fig. 7. Correlation between parameters and PC factors (unrotated) for Period A and Period B.

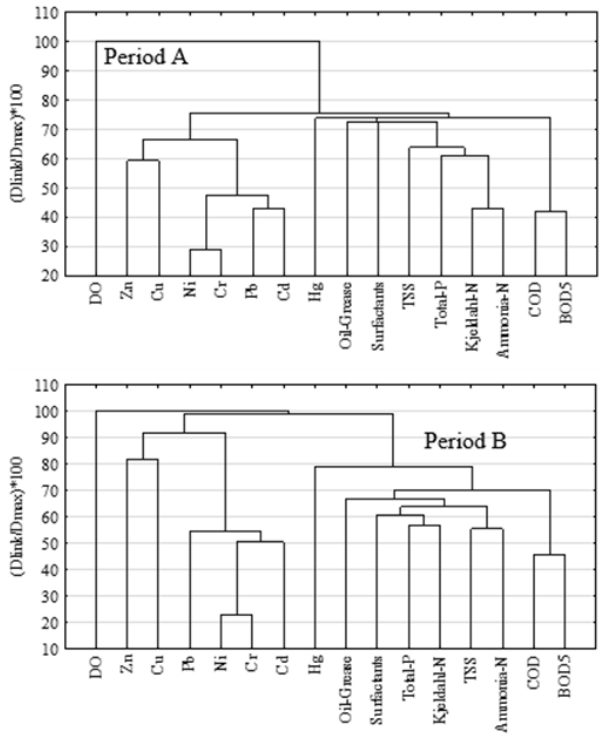

Fig. 8. HCA clustering after PCA of 16 parameters for Period A and Period B.

Conclusions

In this study, water quality dataset obtained from seven different sampling points over 10 years was analysed by multivariate statistical methods. Multivariate analysis was performed by HCA and PCA methods. With HCA analysis, sampling points were grouped into three meaningful clusters with similar characteristics. Moreover, there was no significant change in clusters of Period A and Period B. HCA results indicated that one representative sampling point could be used for water quality monitoring all river network. Additionally, HCA was applied to all water quality parameters to form in meaningful clusters for interpretation relations between parameters. Results indicated that clusters formed during Period A (wet season) have more similar features to whole dataset clusters than the clusters formed for Period B (dry season). PCA was used to obtain a smaller number of variables for the evaluation of surface water quality. Four clusters were obtained for Period A with a total variance of 78.8%, while Period B was grouped into three groups with 68.8% of total variance. The results with the reduced 16 variables in PCA were not consistent with the HCA results of 28 variables. Therefore, another HCA was conducted with 16 significant variables of PCA.

Acknowledgements

The author is thankful to Kayseri Water and Sewage Administration General Directorate (KASKI) for providing water quality data.

References

1. NEDZAREK A., BONISLAWSKA M., TORZ A., GAJEK A., SOCHA M., HARASIMIUK F.B. Water quality in the central reach of the Ina River (Western Pomerania, Poland). Pol. J. Environ. Stud. **24** (1), 207, **2015**.
2. JUNG K.Y., LEE K.L., IM T.H., LEE I.J., KIM S., CHEON S.U., AHN J.M. Evaluation of water quality for the Nakdong River Watershed using multivariate analysis. Environ. Technol. Innov. **5**, 67, **2016**.
3. ALVAREZ-CABRIA M., BARQUIN J., PENAS F.J. Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Sci. Total Environ. **545-546**, 152, **2016**.
4. SHIN P.K.S., FONG K.Y.S. Multiple analysis of marine sediment data. Mar. Pollut. Bull. **39**, 285, **1999**.
5. ZOU S.C., LEE S.C., CHAN C.Y., HO K.F., WANG X. M., CHAN L.Y., ZHANG Z.X. Characterization of ambient volatile organic compounds at a landfill site in Guangzhou, South China. Chemosphere. **51** (9), 1015, **2003**.
6. PANDA U.C., SUNDARAY S.K., RATHA 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. **331**, 434, **2006**.
7. SINGH K.P., MALIK A., MOHAN D., SINHA S. Multivariate statistical techniques for the of spatial and temporal variations in water quality of Gomti River (India)-a case study. Water Res. **38** (18), 3980, **2004**.
8. AZHAR S.C., ARIS A.Z., YUSOFF M.K., RAMLI M.F., JUAHIR H. Classification of river water quality using multivariate analysis. Procedia Environ. Sci. **30**, 79, **2015**.
9. KOSE E., TOKATLI C., CICEK A. Monitoring stream water quality: A statistical evaluation. Pol. J. Environ. Stud. **23** (5) 1637, **2014**.
10. REN J., SHANG Z., TAO L., WANG X. Multivariate analysis and heavy metals pollution evaluation in Yellow River surface sediments. Pol. J. Environ. Stud. **24** (3), 1041, **2015**.
11. LIN W.S., LEE M., HUANG Y.C., DEN W. Identifying water recycling strategy using multivariate statistical analysis for high-tech. Resour. Conserv. Recycl. **94**, 35, **2015**.
12. GÜLER C., KURT M.A., ALPASLAN M., AKBULUT C., Assessment of the impact of anthropogenic and the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) fuzzy clustering, multivariate statistics and GIS techniques. J. Hydrol. **414–415**, 435, **2012**.
13. JIANG Y., GUO H., JIA Y., CAO Y., HU C. Principal component analysis and hierarchical cluster analyses of arsenic groundwater geochemistry in the Hetao Basin, inner Mongolia. Chemie der Erde - Geochemistry. **75** (2), 197, **2015**.
14. MONJEREZI M., VOGT R.D., AAGAARD P., SAKA J.D.K. Hydro-geochemical processes in an area with saline groundwater in lower Shire River Valley, Malawi: An integrated application of hierarchical cluster and principal component analyses. Appl. Geochemistry. **26** (8), 1399, **2011**.
15. ZHAO J., FU G., LEI K., LI Y.W. Multivariate analysis of surface water quality in the Three Gorges Area of China and implications for water management. J. Environ. Sci. **23** (9), 1460, **2011**.
16. OUYANG Y. Evaluation of river water quality monitoring stations by principal component analysis. Water Res. **39**, (12), 2621, **2005**.

17. AJORLO M., ABDULLAH R.B., YUSOFF M.K., HALIM R.A., HANIF A.H.M., WILLMS W.D., EBRAHIMIAN M. Multivariate statistical techniques for the assessment of seasonal variations in surface water quality of pasture ecosystems. Environ. Monit. Assess. **185**, 8649, **2013**.
18. OGWUELEKA T.C. Use of multivariate statistical techniques for the evaluation of temporal and spatial variations in water quality of the Kaduna River, Nigeria. Environ. Monit. Assess. **187**, 137, **2015**.
19. TSI, Turkish Statistical Institute. Land and Climate, Turkey Stat. 2011 Summ. Turkey's Stat. Yearb. Turkey, **2011**.
20. YSKYY, Turkish-Regulation, yüzeysel su kalitesi yönetimi yönetmeliği, Minist. For. Hydraul. Work. Ankara, no. 28483, **2012**.
21. BOYACIOGLU H., BOYACIOGLU H. Investigation of temporal trends in hydrochemical quality of surface water in Western Turkey. Bull. Environ. Contam. Toxicol. **80** (5), 469,**2008**.
22. KOWALKOWSKI T., ZBYTNIEWSKI R., SZPEJNA J., BUSZEWSKI B. Application of chemometrics in river water classification. Water Res. **40** (4), 744, **2006**.
23. GÜLER C., THYNE G.D., MCCRAY J.E., TURNER K.A. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. **10** (4), 455, **2002**.
24. GIBRILLA A., BAM E.K.P., ADOMAKO D., GANYAGLO S., OSAE S., AKITI T.T., KEBEDE S., ACHORIBO E., AHIALEY E., AYANU G., AGYEMAN E.K. Application of water quality index (WQI) and multivariate analysis for groundwater quality assessment of the Birimian and Cape Coast Granitoid Complex: Densu River Basin of Ghana. Water Qual. Expo. Heal. **3**, (2), 63, **2011**.
25. TIRI A., LAHBARI N., BOUDOUKHA A. Assessment of the quality of water by hierarchical cluster and variance analyses of the Koudiat Medouar Watershed, East Algeria. Appl. Water Sci. 10.1007/s13201-014-0261-z, In press, **2016**
26. SHRESTHA S., KAZAMA F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River Basin, Japan. Environ. Model. Softw., **22** (4), 464, **2007**.
27. BELKHIRI L., BOUDOUKHA A., MOUNI L. A multivariate statistical analysis of groundwater chemistry data. Int. J. Environ. Res. **5** (2), 537, **2011**.