Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

ISSN: 0972-6268	Vol. 13	No. 2	pp. 225-234	2014

Original Research Paper

Water Quality Assessment Using Multivariate Statistical Techniques: A Case Study of Yangling Section, Weihe River, China

Xiuquan Xu and Jianen Gao*
Institute of Soil and Water Conservation, CAS & MWR, University of Chinese Academy of Sciences, Yangling-712100, China
*Corresponding Author: Institute of Soil and Water Conservation, CAS & MWR, College of Natural Resources and Environment, College of Water Resources and Architectural Engineering, Institute of Soil and Water Conservation, Northwest A&F University, Yangling-712100, China

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com
Received: 28-8-2013
Accepted: 3-10-2013

Key Words:
Water quality assessment
Multivariate statistical techniques
Weihe river

ABSTRACT

Multivariate statistical techniques, including cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA), were applied for the evaluation of temporal and seasonal variations and interpretation of a complex water quality data set at Yangling Section of Weihe River. Hierarchical cluster analysis grouped 12 months into three clusters, i.e., C1 (relatively highly polluted months), C2 (moderate polluted months) and C3 (less polluted months), based on the similarity of water quality characteristics. Factor analysis/principal component analysis, tested to the data sets of the three groups obtained from cluster analysis, identified 9, 6 and 7 latent factors explaining more than 76, 69 and 62% of the total variance in the data sets of C1, C2 and C3, respectively. The varifactors obtained indicate that parameters responsible for variation are mainly related to temperature and DO (natural), COD_{Mn}, turbidity, NH_4^+, TN, pH and TOC (point source: domestic wastewater) in C1; temperature, DO and EC (natural), COD_{Mn}, TN, pH, and TOC in C2; and temperature, DO and EC (natural), COD_{Mn}, pH and TOC (point source: domestic wastewater and industrial effluents), turbidity and TN (non-point source: agriculture and soil erosion) in C3. However, discriminant analysis showed no significant data reduction, as it used 8 parameters (turbidity, EC, NH_4^+, DO, TN, pH, temperature and TOC) affording more than 81% correct assignations in temporal analysis, while 8 parameters (COD_{Mn}, turbidity, EC, DO, TN, pH, temperature, TOC) affording more than 88% correct assignations in seasonal analysis. Thus, this research illustrated the necessity and usefulness of multivariate statistical techniques for analysis and interpretation of large complex water quality data sets, identification of possible pollution sources/factors and information about variation in water quality for effective river water quality management.

INTRODUCTION

River water is very vulnerable to pollution. The deterioration in water quality could be due to both the natural process, such as variation in precipitation, erosion, weathering of crustal materials, as well as anthropogenic influences viz., urban, industrial and agricultural activities, increasing consumption of water resources especially in the regions with impressive growth of human population and economic development. The quality of rivers at certain places could be reflection of major influences, including the lithology of the basin, atmospheric inputs, climatic conditions and human activities (Simeonov et al. 2003, Shrestha & Kazama 2007, Ellison et al. 2009). Therefore, reliable information on quality is required in order to adopt rationale measures for preventing and controlling the river pollution, and then management of the river basin (Zhang et al. 2011). China has built many water quality automatic monitoring stations all over the country. This results in a huge and complex data matrix comprised of a large number of physical, chemical and biological parameters, which are usually difficult for obtaining meaningful conclusions (Zhao et al. 2011). In the previous studies, the multivariate statistical techniques have been proved to be appropriate tools to solve this problem (Akbal et al. 2011).

The multivariate statistical techniques such as cluster analysis (CA), factor analysis (FA), principal component analysis (PCA) and discriminant analysis (DA) have been widely used as unbiased methods in the interpretation of complex data sets to better understand the water quality and ecological status of the studied water body, identification of possible pollution sources or factors and providing a valuable mean for reliable management of water resources as well as solution to pollution problems. Multivariate statistical techniques have been widely used to evaluate and characterize surface, freshwater and groundwater quality, and proved useful for verifying or evidencing temporal and spatial variations caused by natural and anthropogenic influences linked to seasonality (Liu et al. 2003, Iscen et al. 2008, Sojka et al. 2008, Yerel 2009, Liu et al. 2011).

In the present research, for the first time a large water quality data matrix, during one year (September 2008-August 2009), obtained at the Yangling water quality automatic monitoring station of Weihe River, was subjected to different multivariate statistical techniques (CA, FA/PCA, DA) for drawing meaningful information about quality issue. The aim of the research was to extract information about the similarities or dissimilarities between sampling months, identification of water quality parameters responsible for temporal and seasonal variations in river water quality, the hidden factors explaining the structure of the database, and the influence of possible or potential influences (natural and anthropogenic) on the water quality variables of Weihe River.

MATERIALS AND METHODS

Data source: Weihe River is the largest tributary of the Yellow River, starting from Wushu Mountain in Gansu Province, and going eastward through the Guanzhong Plain (the central of Shanxi Province), then flows into the yellow river at Tongguan County. The river basin has typical continental climate, with the annual mean temperature ranging between 6-13°C. Annual mean precipitation is in the range of 500-800mm, with 60% of precipitation during June-September. The basin is located in the south of the Loess Plateau, belonging to semi-arid climate region. Wei River has a typical characteristic of low discharge and high sediment. Yangling (34°16’ N, 108°04’ E) is located at the centre of Guanzhong Plain (the center of the agriculture and industry in northwest China) (Liu et al. 2007a). The river receives a pollution load from both point (domestic wastewater and industrial effluents) and non-point source (emission of livestock breeding, overuse of fertilizer and soil erosion) (Liu et al. 2007b, Zhang et al. 2007, Geng 2011, Liu et al. 2011).

The Environmental Protection Bureau of Shaanxi Province has built several water quality automatic monitoring stations across the Weihe River Basin, in order to get the new and precise information about the water quality issues just in time. In this research, a one year (September 2008- August 2009) data set of water quality automatic monitoring station at Yangling section was selected. The measured water quality parameters include turbidity, electrical conductivity (EC), ammonia (NH4+), dissolved oxygen (DO), total nitrogen (TN), pH, total phosphorus (TP), temperature (T), chemical oxygen demand (CODMn) and total organic carbon (TOC). The summarized basic statistics of the river water quality dataset is presented in Table 1.

Cluster analysis: Cluster analysis is an unsupervised pattern recognition technique that uncovers intrinsic structure or underlying behaviour of a data set without making a

Table 1: Mean values and standard deviation (S.D.) of different water quality parameters at different months of Wei River (concentration units in mg/L or NTU for turbidity (Turb.); EC in μS/cm and temperature (T) in °C).

Parameters		COD_{Mn}	Turb.	EC	NH_4^+	DO	TN	pH	TP	T	TOC
January	Mean	4.10	30.50	563.00	0.41	6.80	19.76	7.75	0.27	13.60	19.20
	S.D.	2.75	20.85	65.78	0.90	1.07	3.48	0.17	0.86	2.15	1.92
February	Mean	9.05	77.12	661.79	2.40	3.78	24.80	8.00	0.67	12.70	16.80
	S.D.	1.68	27.32	24.84	0.99	1.33	4.51	0.14	0.26	2.28	4.79
March	Mean	8.82	82.28	677.15	4.19	1.44	25.32	8.08	0.55	16.62	34.54
	S.D.	1.98	69.77	27.07	0.99	0.78	2.59	0.07	0.10	3.08	10.36
April	Mean	7.16	34.42	606.34	2.45	2.38	22.47	8.33	0.60	20.23	33.97
	S.D.	2.06	19.83	50.36	2.62	2.74	4.20	0.22	0.14	2.75	8.07
May	Mean	8.93	263.10	439.49	0.47	2.14	13.62	8.65	0.44	20.30	2.55
	S.D.	3.68	273.57	79.87	0.24	0.72	7.36	0.18	0.30	2.16	0.96
June	Mean	7.64	62.42	469.78	0.67	1.22	18.33	8.54	0.33	26.44	3.59
	S.D.	3.13	25.98	176.22	1.89	0.61	14.43	0.31	0.17	2.13	0.76
July	Mean	7.90	296.38	613.88	0.92	6.64	26.93	7.88	1.93	26.40	19.51
	S.D.	5.44	290.21	127.45	0.77	0.77	21.55	0.13	2.53	1.12	19.75
August	Mean	4.14	248.86	529.78	0.66	4.22	4.23	6.64	0.17	22.23	7.43
	S.D.	3.31	352.00	66.18	0.18	1.84	3.76	0.23	0.05	2.88	13.32
September	Mean	14.05	254.46	565.53	1.68	5.95	13.99	6.75	2.49	21.06	4.55
	S.D.	6.75	260.45	56.66	3.10	0.63	9.87	0.15	4.15	1.92	0.72
October	Mean	6.75	255.40	503.66	0.70	5.96	16.96	7.01	0.48	17.66	3.95
	S.D.	2.80	290.08	35.23	2.04	0.58	4.64	0.10	0.24	2.15	1.24
November	Mean	6.34	44.27	536.80	0.21	7.91	7.68	7.07	1.55	14.07	6.03
	S.D.	2.17	18.41	70.72	0.06	0.52	6.87	0.19	2.60	1.66	4.20
December	Mean	4.55	30.62	588.13	0.47	7.03	19.31	7.49	0.33	16.05	19.21
	S.D.	1.61	9.72	30.79	0.30	1.17	1.48	0.09	0.19	3.04	0.98

Table 2: Loadings of variables (10) on significant principal components for C1, C2 and C3.

Variables	VF1	VF 2	VF 3	VF 4
C1 (four significant principal components				
COD_{Mn}	0.094	0.810	0.364	-0.026
Turbidity	-0.150	-0.092	0.802	-0.105
EC	0.395	-0.482	0.333	-0.550
NH_4^+	0.866	0.297	-0.026	0.054
DO	-0.911	-0.082	0.072	-0.050
TN	0.052	0.236	0.756	0.198
pH	0.248	-0.293	0.152	0.823
TP	0.021	0.742	-0.083	-0.132
T	0.818	-0.144	-0.288	-0.030
TOC	0.811	-0.160	0.296	0.121
eigenvalue	3.159	1.731	1.664	1.069
% total variance	31.591	17.313	16.641	10.693
Cumulative % variance	31.591	48.903	65.545	76.238
C2 (three significant principal components				
COD_{Mn}	0.627	0.028	0.517	
Turbidity	0.559	-0.562	0.066	
EC	-0.116	0.409	0.756	
NH_4^+	0.257	0.655	-0.114	
DO	-0.902	-0.114	0.179	
TN	0.140	0.371	-0.836	
pH	0.795	0.188	-0.366	
TP	-0.175	-0.145	0.412	
T	0.853	-0.164	-0.235	
TOC	-0.101	0.907	-0.039	
Eigenvalue	3.108	1.970	1.947	
% total variance	30.118	19.701	19.471	
Cumulative % variance	30.180	49.881	69.352	
C3 (four significant principal components				
COD_{Mn}	0.231	-0.08	-0.236	0.634
Turbidity	-0.158	0.118	0.102	0.769
EC	0.681	0.244	0.449	-0.143
NH_4^+	0.259	0.286	-0.197	-0.202
DO	0.788	0.143	0.155	0.298
TN	-0.081	0.816	0.118	0.3
pH	-0.807	0.142	0.131	0.101
TP	0.109	0.856	0.001	-0.12
T	-0.041	0.124	0.75	0.092
TOC	0.138	-0.105	0.731	-0.149
Eigenvalue	1.922	1.627	1.459	1.289
% total variance	19.216	16.27	14.587	12.889
Cumulative % variance	19.216	35.486	50.073	62.962

priori assumption about the data, in order to classify the objects of the system into categories or clusters based on their nearness or similarity (Vega et al. 1998). Hierarchical clustering is the most common approach in which clusters are formed sequentially, by starting with the most similar pair of objects and forming higher clusters step by step. The Euclidean distance usually gives the similarity between two samples and a 'distance' can be represented by the 'difference' between analytical values from both the samples (Zhang et al. 2011). Hierarchical agglomerative CA was performed on the normalized data set by means of the Ward's method, using Euclidean distance as a measure of similarity. This method uses the analysis of variance approach to evaluate the distance between clusters, attempting to minimize the sum of squares of any two clusters that can be formed at each step. Cluster analysis was applied to the river water quality data set with a view to group the similar sampling months and in the resulted dendrogram. The linkage distance is reported as Dlink/Dmax, which represent the quotient between the linkage distances for a particular case divided by the maximal distance, multiplied by 100 as a way to standardize the linkage distance (Alberto et al. 2001).

Principal component analysis/factor analysis: PCA technique extracts the eigen values and eigen vectors from the covariance matrix of original variables. The PCs are the uncorrelated variables, obtained by multiplying the original correlated variables with the eigen vector, which is list of coefficients (loadings or weightings). Thus, the PCs are weighted linear combinations of the original variables. PC provides information on the most meaningful parameters, which describe whole data set affording data reduction with minimum loss of original information (Iscen et al. 2008). It is a powerful technique for pattern recognition that attempts to explain the variance of a large set of inter-correlated variables and transforming into a smaller set of independent (uncorrelated) variables (principal components).

Factor analysis further reduces the contribution of less significant variables obtained from PCA and the new group of variables known as varifactors (VFs) is extracted by rotating the axis defined by PCA. A VF can include unobservable, hypothetical, latent variables, while a PC is a linear combination of observable water-quality variables (Liu et al. 2003). PCA of the normalized parameters (water-quality data set) was performed to contribution of variables with minor significance, and these PCs were subjected to varimax rotation (raw) generating VFs.

Discriminant analysis: Discriminating analysis is used to determine the variables, which discriminate between two or more naturally occurring groups. It operates on raw data and the technique constructs a discriminant function for each group as in eq. (1):

$$f(G_i) = k_i ij + \sum_{j=1}^{n} w_{ij} p_{ij} \qquad ...(1)$$

Where i is the number of groups (G), k_i is the constant inherent to each group, n is the number of parameters used to classify a set of data into a given group, w_i is the weight coefficient, assigned by DA to a given selected parameter (p_j).

Data treatment: In this case study, three groups for temporal evaluations (three monitoring time regions, obtained by

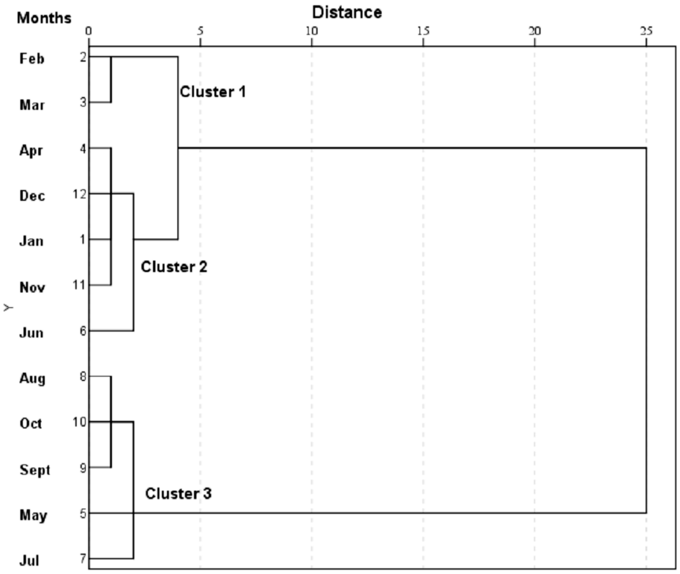

Fig. 1: Dendrogram showing clustering of sampling months according to surface water quality at Yangling section of Weihe River according to Ward's method using squared Euclidean distance.

CA) and four for seasonal (four seasons) ones have been selected, and the number of analytical parameters used to assign a measure from a monitoring time into a group (monitoring time or month). Discriminant analysis was applied to raw data by using three modes, viz. the standard, forward stepwise and backward stepwise modes, to construct discriminant functions to evaluate both the temporal and seasonal variations in river water quality. The temporal and season were the grouping (dependent) variables, while all the measured parameters constituted the independent variables.

RESULTS AND DISCUSSION

Monthly similarity: Cluster analysis was used to detect the similarity groups or periods between the sampling months. It yielded a dendrogram (Fig. 1) grouping all 12 months into three statistically clusters or three periods at $(D_{link}/D_{max}) \times 100 < 20$. The cluster 1 (February and March) corresponds to highly polluted months. In cluster 1, these months correspond to low discharge, and river water temperature, then the pollutants concentration are relatively high and situation is severe. Cluster 2 (November-January, April and June) corresponds to moderately polluted months. These months correspond to high precipitation and discharge and high river water temperature, and as a result, cluster 2 group suggests the dilution effect and self-purification and assimilative capacity of the river. Cluster 3 (May and July-October) corresponds to relatively less polluted months, in these months hydrological factors are moderate compared to the two above. The results indicate that CA technique is useful in offering reliable classification of surface water quality in the whole region and will make it possible to design a future temporal sampling strategy in an optimal manner, which could reduce the number of sampling times and associated costs. There are other reports (Shrestha & Kazama 2007), where similar approach has successfully applied in water quality monitoring programs.

Temporal variations in river water quality: Principal component analysis/factor analysis was performed on the data set (10 variables) separately for the three different time regions, viz. C1, C2 and C3, as delineated by CA techniques, to compare the compositional pattern between analysed water samples and identify the factors influencing each one. The input data matrices [variables*cases] for PCA/FA were [10*244] for C1, [10*705] for C2 and [10*794] for C3. PCA of the three data sets yielded 4 PCs for C1 and C3 and 3 PCs for C2 with eigen value >1, explaining 76.24, 62.96 and 69.35% of the total variance in respective water quality data sets (Table 2). An eigen value gives a measure of the significance of the factor: the factor with the highest eigen value is the most significant. Eigen values of 1.0 or greater are considered significant (Shrestha & Kazama 2007). Unequal numbers of VFs were obtained for three time regions through FA performed on the PCs. Corresponding VFs, variable loading and explained variance are presented in Table 2. Liu et al. (2011) classified the factor loading as 'strong', 'moderate' and 'weak', corresponding to absolute loading values of >0.75, 0.75-0.50 and 0.50-0.30, respectively.

For the data set pertaining to C1 (relatively highly polluted months), among four VFs, VF1, explaining 31.591%

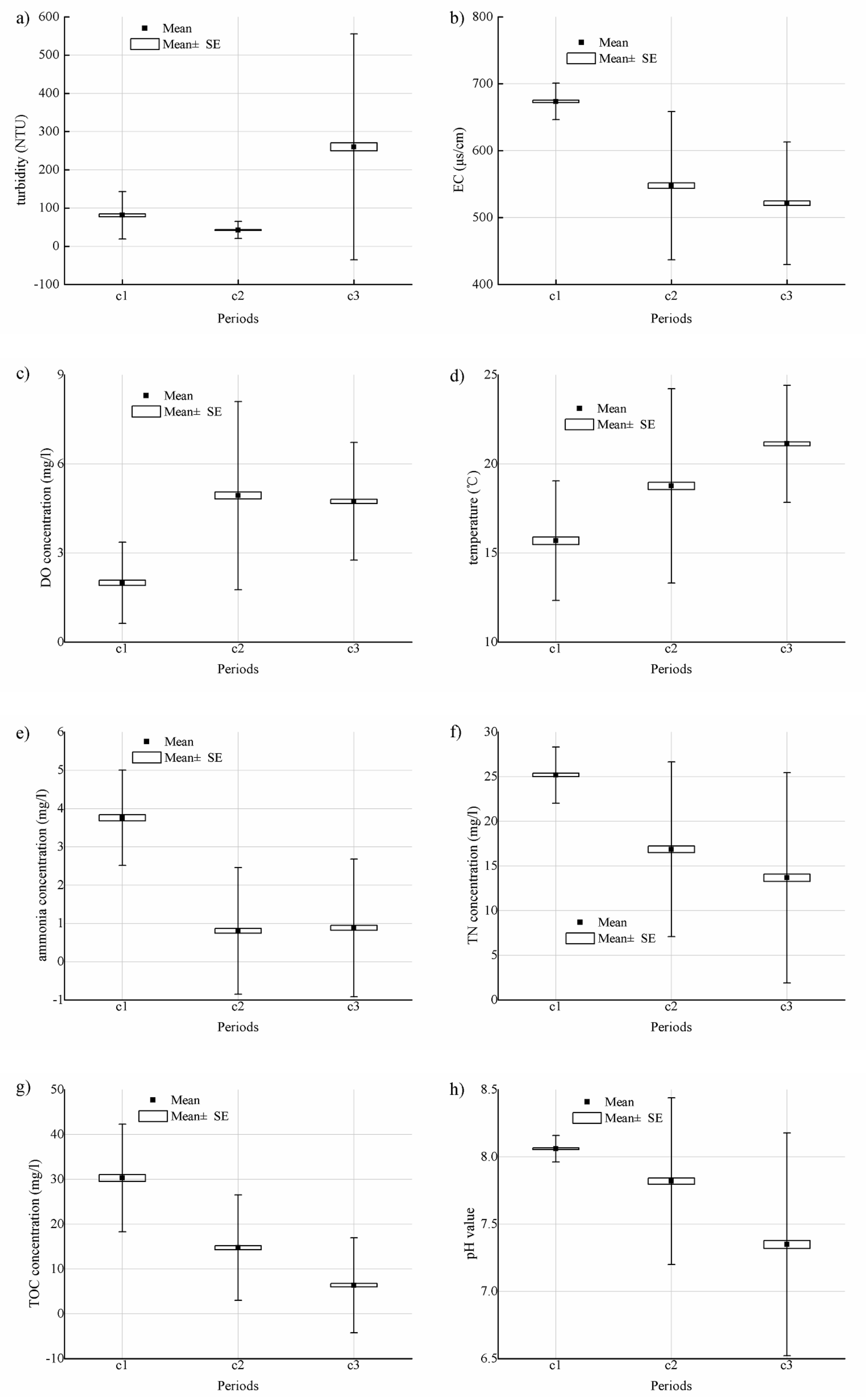

Fig. 2: Temporal variations: (a) turbidity, (b) EC, (c) DO, (d) temperature, (e) NH_4^+, (f) TN, (g) TOC and (h) pH value in surface water quality at Yangling section of Weihe River (error bar means Mean±SD).

Table 3: Classification matrix for discriminant analysis of temporal variation in water quality.

Months	% correct	Assigned by DA		
		C1	C2	C3
Standard DA mode				
C1	98.8	241	51	4
C2	78.4	3	553	165
C3	78.7	0	101	625
Total	81.4	244	705	794
Forward DA mode				
C1	98.8	241	51	4
C2	78.4	3	553	165
C3	78.7	0	101	625
Total	81.4	244	705	794
Backward DA mode				
C1	98.8	241	51	4
C2	80.7	3	569	183
C3	76.4	0	85	607
Total	81.3	244	705	794

of total variance, has strong positive loading on NH_4^+, temperature and TOC and strong negative loading on DO. The inverse relationship between temperature and DO is a natural process because warmer water gets saturated more easily with oxygen and then it could hold less dissolved oxygen. VF2, explaining 17.313% of the total variance, has strong positive loading on COD_{Mn} and TP. VF1 and VF2 may represent the organic pollution from industrial and domestic waste. VF3, explaining 16.641% of the total variance, has strong positive loading for turbidity and TN. The relationship between turbidity and TN may represent a certain amount of non-point source pollution. VF4, explaining 10.693% of total variance, has strong positive loading on pH value. In February and March, the non-point source pollution caused by precipitation and runoff is limited, so the organic pollution was mainly caused by anthropogenic activities, such as discharges from wastewater treatment plants, domestic wasterwater and industrial effluents.

For the data set representing C2 (moderate polluted months), among three VFs, VF1, explaining 30.118% of total variance, has strong positive loading on temperature and strong negative loading on DO, representing the seasonal impact, which is the same with the inverse relationship above. VF2, explaining 19.701% of the total variance, has strong positive loading on TOC. VF3, explaining 19.471% of the total variance, has strong positive loading for EC and strong negative loading on TN. Variation in EC may show the anti-dilution effect and also a natural process. The strong positive loading for TOC and negative for TN may be due to the increasing influence of industrial effluents and weakening of non-point source pollution.

Lastly, for the data set pertaining to water quality in C3 (less polluted months), among four VFs, VF1, explaining 19.216% of total variance, has strong positive loading on dissolved oxygen, and moderate positive loading on EC, and strong negative loading for pH. This factor may be due to aerobic conditions and hence strengthened dilution effect and water self-purification effect in the river. VF2, explaining 16.27% of the total variance, has strong positive loading on TN and TP. This factor represents the contribution of non-point source pollution from upland area. In these areas, farmers use the nitrogenous and phosphorus fertilizers, and the river receive total nitrogen and total phosphorus from runoff and sediment caused by precipitation. VF3, explaining 14.587% of the total variance, has strong positive loading for temperature and TOC. The variation of temperature can be attributed to seasonal changes. VF4, explaining 12.889% of total variance, has strong positive loading on turbidity and moderate positive loading on COD_{Mn}, and this result may explain the non-point source pollution as the same with C1. The strong positive loading in TOC, and moderate positive loading in COD_{Mn} represent the organic pollution from domestic waste. As a result, pollution from neither point source nor non-point sources should not be omitted.

Temporal variations in water quality were future evaluated through DA. Temporal DA was performed on raw data after dividing the whole data set into three periods, C1, C2 and C3. Discriminant functions (DFs) and classification matrices (CMs) obtained from the standard, forward stepwise and backward stepwise modes of DA are given in Tables 3 and 4. In forward stepwise mode, variables are included step-by-step beginning with the more significant until no significant changes are obtained, while in backward stepwise mode, variables are removed step-by-step beginning less significant until no significant changes are obtained. Temporal DA was performed with the same raw data set comprising 10 parameters after grouping into three major classes of C1, C2 and C3 as obtained through CA. The time regions (clustered) were the grouping (dependent) variable, while all the measured parameters constituted the independent variables. Both the standard and forward stepwise mode DFs using the total 10 parameters yielded the corresponding CMs assigning more than 81% cases correctly (Tables 3 and 4). However, the backward stepwise mode DA gave CMs with > 80% correct assignations using only 8 discriminant parameters (Tables 3 and 4) with little difference in match for each period compared with the other two modes. Backward stepwise DA shows that turbidity, EC, NH_4^+, DO, TN, pH, temperature and TOC are the discriminating parameters.

Box and whisker plots of discriminating parameters identified by temporal DA were constructed to evaluate different patterns associated with temporal variations in river wa-

Table 4: Classification function for discriminant analysis of temporal variations in water quality.

parameters	Standard DA mode			Forward stepwise DA mode			Backward stepwise DA mode		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
COD_{Mn}	-0.177	-0.280	-.190	-0.177	-0.280	-0.190			
Turb.	0.005	0.003	0.008	0.005	0.003	0.008	0.005	0.002	0.007
EC	0.124	0.102	0.099	0.124	0.102	0.099	0.120	0.099	0.097
NH_4^+	3.258	2.599	2.674	3.258	2.599	2.674	3.108	2.427	2.556
DO	5.304	6.468	6.355	5.304	6.468	6.355	5.241	6.408	6.314
TN	-0.264	-0.364	-0.395	-0.264	-0.364	-0.395	-0.294	-0.378	-0.405
PH	32.941	33.123	31.378	32.941	33.123	31.378	32.900	32.942	31.258
TP	-0.654	-0.440	-0.313	-0.654	-0.440	-0.313			
T	0.647	1.224	1.412	0.647	1.224	1.412	0.638	1.214	1.405
TOC	-.0289	-0.323	-0.369	0.289	-0.323	-0.369	-0.260	-0.294	-0.349
constant	-183.641	-180.772	-169.614	-183.641	-180.772	-169.614	-182.837	-179.903	-169.204

Table 5: Classification matrix for discriminant analysis of seasonal variation in water quality.

Season	% correct	Season assigned by DA			
		Spring	Summer	Autumn	Winter
Standard DA mode					
Spring	89.0	259	38	0	39
Summer	77.1	31	377	0	0
Autumn	98.7	0	68	452	24
Winter	87.5	1	6	6	442
Total	87.8	291	489	458	505
Forward DA mode					
Spring	89.3	260	38	0	40
Summer	77.7	31	380	0	0
Autumn	99.1	0	64	454	24
Winter	87.3	0	7	4	441
Total	88.1	291	489	458	505
Backward DA mode					
Spring	90.4	263	37	0	39
Summer	77.7	27	380	0	0
Autumn	99.1	0	65	454	24
Winter	87.5	1	7	4	442
Total	88.3	291	489	458	505

ter quality (Fig. 2). The average turbidity (Fig. 2-a) is highest in C3 compared to C1 and C2, and this fact might have been due to the great amount of suspended solids caused by high discharge in C3, which could be proved by the study about the relationship between discharge and suspended solids in the Yellow river, China (UNEP GEMS/WATER PROGRAMME 2008). The average concentration of EC (Fig. 2-b) is highest in C3 compared to C1 and C2, which represents a dilution effect. The average concentration of DO (Fig. 2-c) is observed to be highest in C2 and lowest in C3, while the average temperature (Fig. 2-d) highest in C3 and lowest in C1, which is due to seasonality effect. There is an inverse relationship between DO and temperature. This difference could be explained that the inverse relationship between both is a nature-oriented process because warmer water gets saturated with oxygen more easily and hence could hold less dissolved oxygen, whereas in this study it is a pollutant-oriented process because the river water especially in C1 has a high concentration of organic pollutants which could result in decrease of solubility of oxygen. The average concentration of NH_4^+ (Fig. 2-e) is highest in C1 and lowest in C2. The average concentrations of TN (Fig. 2-f) and TOC (Fig. 2-g) show the same pattern. The average pH values (Fig. 2-h) are all above 7, showing that the surface water is alkaline. A clear inverse pattern between organic pollutants and pH is observed, to which a rational explanation would be that the organic pollutants lead to anaerobic conditions in the river water, and then result in formation of ammonia and organic acids, which could cause a decrease of water pH. The same trend have been observed in other studies (Wang et al. 2011).

Seasonal variations in water quality: Seasonal variations in water quality were future evaluated through DA. Seasonal DA was performed on raw data after dividing the whole data set into four season groups (Spring: March-May; Summer: June-August; Autumn: September-November and Winter: December-February). Discriminant functions (DFs) and classification matrices (CMs) obtained from the standard, forward stepwise and backward stepwise modes of DA are given in Tables 5 and 6. The standard DA mode and forward stepwise DA mode-constructed DFs, including 10 and 9 parameters, respectively, are given in Tables 5 and 6. The standard mode of DFs using 10 discriminant variables yielded the corresponding CMs assigning 87.8% of the cases correctly (Tables 5 and 6). The forward stepwise mode of DFs using 9 discriminant variables yielded the corresponding CMs assigning 88.1% of the cases correctly (Tables 5 and 6). However, in backward stepwise mode, DA gave CMs with 88.3% correct assignations using only 8 discriminant parameters (Tables 5 and 6) with little difference in match for each season compared with the other two modes. Thus, the seasonal DA results suggest that COD_{Mn}, turbidity, EC,

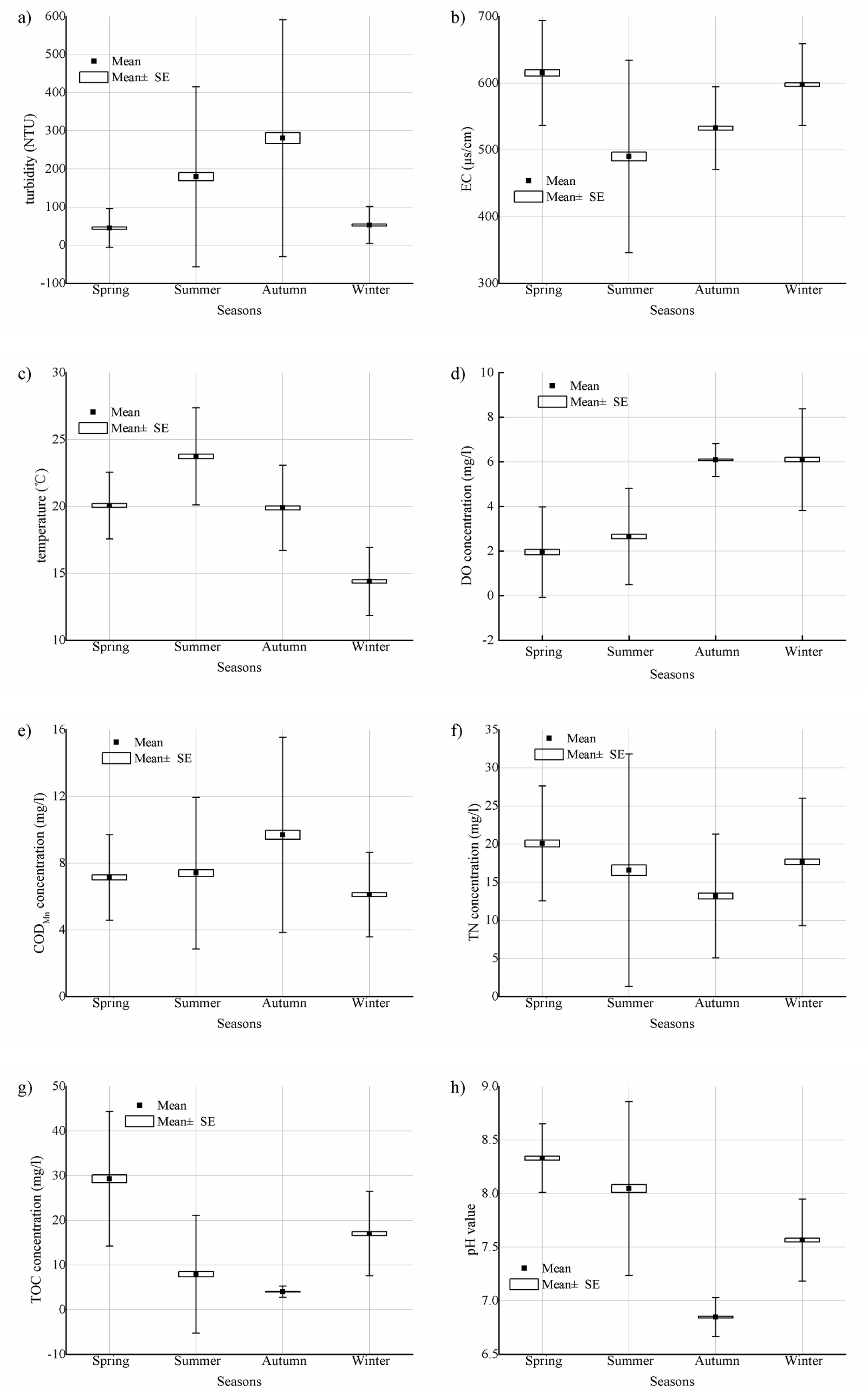

Fig. 3: Seasonal variations: (a) turbidity, (b) EC, (c) temperature, (d) DO, (e) COD_{Mn}, (f) TN, (g) TOC, and (h) pH value in surface water quality at Yangling section of Wei River (error bar means Mean±SD).

Table 6: Classification function for discriminant analysis of seasonal variations in water quality.

Parameters	Standard DA mode				Forward stepwise DA mode				backward stepwise DA mode			
	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
COD_{Mn}	-1.416	-1.185	-0.879	-1.317	-1.417	-1.187	-0.876	-1.313	-1.228	-1.024	-0.694	-1.139
Turbidity	-0.003	-0.001	0.003	-0.002	-0.003	-0.001	0.003	-0.002	-0.005	-0.002	0.002	-0.003
EC	0.125	0.111	0.113	0.129	0.125	0.112	0.113	0.128	0.126	0.112	0.114	0.129
NH_4^+	2.931	2.528	2.831	2.707	2.935	2.534	2.823	2.692				
DO	4.124	4.030	4.874	5.081	4.129	4.038	4.865	5.064	3.194	3.231	3.966	4.206
TN	-0.813	-0.741	-0.646	-0.698	-0.812	-0.738	-0.649	-0.705	-0.698	-0.640	-0.540	-0.601
pH	52.717	49.059	43.395	49.389	52.706	49.042	43.414	49.429	51.016	47.582	41.788	47.879
TP	0.037	0.059	-0.067	-0.136								
T	1.443	1.866	1.503	0.800	1.444	1.867	1.503	0.799	1.285	1.730	1.350	0.653
TOC	0.308	0.139	0.008	0.139	0.306	0.137	0.010	0144	0.409	0.226	0.110	0.239
Constant	-273.26	-244.44	-203.29	-240.29	-273.26	-244.44	-203.28	-240.26	-263.13	-236.88	-193.90	-231.74

DO, TN, pH, temperature and TOC are the most significant parameters to discriminate between the four seasons, which means that these parameters account for most of the expected seasonal variations in the water quality (Tables 5 and 6).

As identified by DA, box and Whisker plots of the selected parameters showing seasonal trends are given in Fig. 3. The average turbidity (Fig. 3-a) increases from spring to autumn with decrease in winter, which may be due to variation in suspended solids caused by discharge. The average concentration of EC (Fig. 3-b) is relatively lower in summer and autumn compared to spring and winter, showing a dilution effect. The average temperature (Fig. 3-c) reveals a clearcut seasonal effect. The average concentration of DO (Fig. 3-d) shows two distinct groups, with low in spring and summer and high in august and winter. As discussed above, in natural condition there would be an inverse relationship between temperature and DO, so the low DO concentration in spring might be reflection of anaerobic environment caused by organic pollution. The average concentration of COD_{Mn} (Fig. 3-e) increases from spring to autumn with decrease in winter, and it is the same case with turbidity (Fig. 3-a). The average concentrations of TN (Fig. 3-f), TOC (Fig. 3-g) and pH (Fig. 3-h) show the same trend with decrease from spring to autumn and increase in winter. The low pH value in autumn might be due to high concentration of organic pollutant, e.g. the highest concentration of COD_{Mn} in spring (Fig. 3a). The same seasonal variation between organic pollutants and pH have been reported (Vega et al. 1998, Shrestha & Kazama 2007, Sojka et al. 2008).

CONCLUSIONS

Water quality automatic monitoring programs generate complex multidimensional data which need multivariate statistical treatment for their analysis in order to get the underlying information. In this case study, different multivariate statistical techniques (CA, FA/PCA, DA) were used to estimate and evaluate the different temporal and seasonal variations in surface water quality at Yangling section of Weihe River, China. Hierarchical cluster analysis (CA) successfully grouped 12 sampling months into three clusters based on the similar water quality characteristics, corresponding to C1 (highly polluted time region), C2 (moderately polluted) and C3 (less polluted), respectively. The information obtained by CA is helpful to make an optimal sampling strategy by reduction of sampling times and hence the associated costs. The factor analysis/principal component analysis (FA/PCA) could help extract and identify the factors and related sources responsible for variations in river water quality at three different sampling times. Varifactors obtained from factor analysis indicate that parameters responsible for water quality variation are mainly related to temperature and DO (natural), COD_{Mn}, turbidity, NH_4^+, TN, pH and TOC (point source: domestic wastewater) in C1 (HP); temperature, DO and EC (natural), COD_{Mn}, TN, pH, and TOC in C2 (MP); and temperature, DO and EC (natural), COD_{Mn}, pH and TOC (point source: domestic wastewater and industrial effluents), turbidity and TN (non-point source: agriculture and soil erosion) in C3 (LP). Discriminant analysis (DA) helped in discriminatory and identifying significant parameters for discrimination among temporal and seasonal groups, and it gave not significant but encouraging results. It used 8 parameters (turbidity, EC, NH_4^+, DO, TN, pH, temperature, TOC) affording more than 81% correct assignations in temporal analysis, and 8 parameters (COD_{Mn}, turbidity, EC, DO, TN, pH, temperature, TOC) affording more than 88% correct assignations in seasonal analysis. Thus, the multivariate statistical techniques proved to be an excellent exploratory and useful tool in analysis of complex water quality data set and in understanding their temporal and seasonal variations.

ACKNOWLEDGMENT

This work was supported by the National Technology Support Project of China (2011BAD31B05), the Natural Science Foundation of China (41371276), the Subject of National Science and Technology Major Project of China (2009ZX07212-002-003-02) and the Knowledge Innovation Project of Institute of Soil and Water Conservation, CAS & MWR (Soil and Water Conservation Project) (A315021304).

REFERENCES

Akbal, F., Gurel, L., Bahadir, T., Guler, I., Bakan, G. and Buyukgungor, H. 2011. Multivariate statistical techniques for the assessment of surface water quality at the Mid-Black Sea Coast of Turkey. Water Air Soil Poll., 216: 21-37.

Alberto, W.D., Del Pilar, D.M., Valeria, A.M., Fabiana, P.S., Cecilia, H.A. and Angeles, B.M. 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River basin (Cordoba-Argentina). Water Res. 35: 2881-2894.

Ellison, C.A., Skinner, Q.D. and Hicks, L.S. 2009. Assessment of best-management practice effects on rangeland stream water quality using multivariate statistical techniques. Rangeland Ecol. Manag., 62: 371-386.

Geng, Y.N. 2011. Water quality assessment of Baoji Reach of Weihe river based on fuzzy conposite index method. Yellow River, 33: 36-37. (in Chinese)

Han, T., Li, H.E. and Li, J.K. 2004. Analysis of nitrogen pollution above lintong sectong of Weihe river. Yellow River, 26: 22-23. (in Chinese)

Iscen, C.F., Emiroglu, O., Ilhan, S., Arslan, N., Yilmaz, V. and Ahiska, S. 2008. Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. Environ. Monit. Assess., 144: 269-276.

Liu, C.W., Lin, K.H. and Kuo, Y.M. 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Sci. Total Environ., 313: 77-89.

Liu, Y., Hu, A.Y. and Deng, Y.Z. 2007a. Temporal and spatial evolution characters of water quality in Weihe river basin in Shaanxi Province. Water Resource Protection, 23: 11-13. (in Chinese).

Liu, Y. and Hu, A.Y. 2007b. Cause of formation of water problems in Weihe river basin and countermeasures. Water Resource Protection. 23: 17-21. (in Chinese).

Li, J.K., Li, H.E., Dong W., Qin, Y.M., Huang, C.J. and Du, G.F. 2011. Monitoring and load estimation of non-point source pollution on typical tributaries in the guanzhong reach of the Weihe river. Acta Scientiae Circumstantiae, 31: 1470-1478. (in Chinese).

Liu, W.C., Yu, H.L. and Chung, C.E. 2011. Assessment of water quality in a subtropical alpine lake using multivariate statistical techniques and geostatistical mapping: A case study. Int. J. Env. Res. Pub. Hlth, 8(4): 1126-1140.

UNEP GEMS/WATER PROGRAMME 2008. Water Quality for Ecosystem and Human Health, Second Edition, 3rd ed. United Nations Environment Programme Global Environment Monitoring System (GEM)/Water Programme: Burlington, Ontario, Candna, pp. 13.

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

Wang, T.R., Sun, G.N. and Liu, S. Y. 2011. Relationship between spatiotemporal variation of water pollution and runoff volume of mainstream sectioning of the Weihe river in Shaanxi province. Arid Zone Research, 28: 609-615. (in Chinese).

Shrestha, S. and Kazama, F. 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environ Modell Softw., 22: 464-475.

Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. and Kouimtzis, T. 2003. Assessment of the surface water quality in Northern Greece. Water Res., 37: 4119-4124.

Sojka, M., Siepak, M., Ziola, A., Frankowski, M., Murat-Blazejewska, S. and Siepak, J. 2008. Application of multivariate statistical techniques to evaluation of water quality in the Mala Welna River (Western Poland). Environ. Monit. Assess., 147: 159-170.

Yerel, S. 2009. Assessment of surface water quality using multivariate statistical analysis techniques: A case study from Tahtali dam, Turkey. Asian J. Chem., 21: 4054-4062.

Zhang, I., Zhao, C.C., Lin, J.H., Wang, X.C. and Kusuda T. 2007. Application of a distributed hydrological model for characterization of Weihe river basin. Journal of Xi'an University of Arch. & Tech. (Natural Science Edition). 39: 61-65. (in Chinese).

Zhang, X.A., Wang, Q.S., Liu, Y.F., Wu, J. and Yu, M.A. 2011. Application of multivariate statistical techniques in the assessment of water quality in the southwest new territories and Kowloon, Hong Kong. Environ. Monit. Assess., 173: 17-27.

Zhao, G., Gao, J., Tian, P., Tian, K. and Ni, G. 2011. Spatial-temporal characteristics of surface water quality in the Taihu Basin, China. Environ. Earth Sci., 64: 809-819.