

Evaluation of river water quality variations using multivariate statistical techniques

Sava River (Croatia): a case study

Andrea Marinović Ruždjak · Domagoj Ruždjak

Received: 18 January 2014 / Accepted: 4 March 2015 / Published online: 29 March 2015
© Springer International Publishing Switzerland 2015

Abstract For the evaluation of seasonal and spatial variations and the interpretation of a large and complex water quality dataset obtained during a 7-year monitoring program of the Sava River in Croatia, different multivariate statistical techniques were applied in this study. Basic statistical properties and correlations of 18 water quality parameters (variables) measured at 18 sampling sites (a total of 56,952 values) were examined. Correlations between air temperature and some water quality parameters were found in agreement with the previous studies of relationship between climatic and hydrological parameters. Principal component analysis (PCA) was used to explore the most important factors determining the spatiotemporal dynamics of the Sava River. PCA has determined a reduced number of seven principal components that explain over 75 % of the data set variance. The results revealed that parameters related to temperature and organic pollutants (COD_{Mn} and TSS) were the most important parameters contributing to water quality variation. PCA analysis of seasonal subsets confirmed

this result and showed that the importance of parameters is changing from season to season. PCA of the four seasonal data subsets yielded six PCs with eigenvalues greater than one explaining 73.6 % (spring), 71.4 % (summer), 70.3 % (autumn), and 71.3 % (winter) of the total variance. To check the influence of the outliers in the data set whose distribution strongly deviates from the normal one, in addition to standard principal component analysis algorithm, two robust estimates of covariance matrix were calculated and subjected to PCA. PCA in both cases yielded seven principal components explaining 75 % of the total variance, and the results do not differ significantly from the results obtained by the standard PCA algorithm. With the implementation of robust PCA algorithm, it is demonstrated that the usage of standard algorithm is justified for data sets with small numbers of missing data, nondetects, and outliers (less than 4 %). The clustering procedure highlighted four different groups in which the sampling sites have similar characteristics and pollution levels. The first and the second group correspond to relatively low and moderately polluted sites while stations which are located in the middle of the river belong to the third and fourth group and correspond to highly and moderately polluted sites.

A. Marinović Ruždjak (✉)
Croatian Waters, Central Water Management Laboratory,
Ulica grada Vukovara 220, 10000 Zagreb, Croatia
e-mail: andrea.marinovic@zg.t-com.hr

D. Ruždjak
Hvar Observatory, University of Zagreb,
Faculty of Geodesy, Kačićeva 26,
10000 Zagreb, Croatia

Keywords Principal component analysis (PCA) · Cluster analysis (CA) · Robust PCA · Surface water quality

Introduction

Surface water quality is very important and very sensitive issue. It is determined by natural processes such as dissolution of geological deposits, biological degradation of organic matter, and atmospheric deposition. In addition to natural sources, urban and industrial development, farming, building of dams, and diversion of flow and other human activities within the aquatic environment alter the physical and chemical composition of water (Meybeck 1998; Malmqvist and Rundle 2002; Mayer et al. 2002). Since degradation of water quality can result in altered species diversity, decrease the overall health of aquatic ecosystem, and cause a serious harm to human health and the environment, it is necessary to monitor the water quality regularly. Long-term water quality monitoring programs (Dixon and Chiswell 1996) often provide large sets of data which are difficult to interpret. In attempt to draw meaningful conclusions and achieve better understanding of water quality, different multivariate techniques, such as principal component analysis (PCA), factor analysis (FA), or cluster analysis (CA), have been applied to these complex data matrices (see, e.g., Vega et al. 1998; Simeonov et al. 2003; Bouza-Deano et al. 2008; Li et al. 2011; Pinto and Maheshwari 2011; Olsen et al. 2012; Wang et al. 2012; Mei et al. 2014).

PCA is widely used technique in science. It was often applied to exploratory data analysis in water quality research as has been recently summarized by Olsen et al. (2012). PCA is the simple non-parametric method which can be used to reduce a complex data set to a lower dimension and capture sometimes hidden, underlying patterns present within the original data (Wold et al. 1987). It is optimal in terms of mean squared error, and the model parameters can be computed directly from the data. Despite these attractive features, PCA models have several shortcomings. First, it is not obvious how to deal properly with incomplete data set (what is often the case). Second, the standard PCA algorithm is based on the assumption that data are not spoiled by outliers. In case of outliers, robust version of PCA has to be developed (Ruymgaart 1981; Maronna et al. 2006; Stanimirova et al. 2007). To account for outliers, the most commonly used approach is to replace the

standard estimation of the covariance matrix with a robust one (Campbell 1980; Croux and Haesbroeck 2000; Hubert et al. 2002).

The main goal of this study is to determine the state of the Sava River in the transition period between the end of the war in Croatia and the adjustment of its legislation to the European standards. Therefore, the results of this study can be used as 'baseline' condition when examining the impacts of those changes. For this purpose, different multivariate statistical approaches were applied to the data matrix obtained during the 7-year water quality monitoring of the Sava River in Croatia with the goal to detect and quantify the water quality trends. Pearson correlation coefficient, PCA, and CA were used to obtain the information about the latent factors explaining the data set, to describe the seasonal and spatial variations, the influence of possible sources (natural or anthropogenic) on the parameters, and the similarities between sampling sites. Robust PCA algorithm was used to check the influence of the outliers in the data set on the results. All mathematical and statistical computations were made using MATLAB version 7.7.0.471 (R2008B).

Materials and methods

Study area

Water samples used in this work were collected from the Sava River in Croatia. The river is located in Southeast Europe. It has its source in Slovenia, flows through Croatia, Bosnia and Herzegovina, and on to Serbia where it joins, as side tributary, river Danube in Belgrade. It is suitable for navigation downstream of Sisak, Croatia. The river is 945 km long; in Croatia, it flows in the length of 510 km and for the most part constitutes the border with Bosnia and Herzegovina (313 km). The average flow of the Sava River at its entry into Croatia (Jesenice) and at its exit from Croatia amounts to 300 and 1200 m³/s, respectively. The main tributaries in this segment of the basin are Kupa, Una, Vrbas, Sutla, Krapina, Lonja, Ilova, Orpljava, and Bosut. Major towns along the river that have considerable industrial capacities are Zagreb, Sisak, Slavonski Brod, and Županja. The main sources of pollution

come from untreated wastewaters discharged from municipalities and industries along the Sava River as well as the tributaries.

In its middle course, downstream of Zagreb, Sava becomes a very large lowland river; it is unique example of a river with flood plains that are still intact and support flood alleviation and biodiversity. It has been considered as one of the key areas in the Pan European Biological and Landscape Diversity Strategy (PEBLDS) of the Council of Europe. The Lonjsko Polje Nature Park and Special ornithological reserve Crna Mlaka (Crna Mlaka Fish-ponds) have been included on the Ramsar List of Wetlands of International Importance and on the list of internationally Important Bird Areas. The Lonsko Polje has also been included in the ecological network in its entirety as an area important for the conservation of species and habitats. In the Sava River Basin in Croatia, almost entire public water supply has been based upon groundwater exploitation. Since the water of the Sava River is in the hydraulic connection with groundwater, the quality and the quantity of the Sava River are very important.

Monitoring sites

In Croatia, the water quality monitoring network is operated primarily by *Croatian Waters*. The Sava River monitoring system covers 18 sample sites

(Fig. 1) where the water quality is tested twice per month. Water quality monitoring stations are located upstream and downstream of the tributaries, developed areas and inputs from drains. The first and the last station (S1, S18) are located just downstream/upstream of the international border with Slovenia/Serbia. Stations S6, S8, S10, and S14 are located upstream while the stations S7, S9, S11, and S15 downstream of the right tributaries Kupa, Una, Vrbas, and Bosna, respectively. Stations S6 and S7 are also covering an area of Sisak. Stations S2, S12, and S16 are located upstream while the stations S3, S4, S5, S13, and S17 downstream of Zagreb, Slavonski Brod and Županja. Station S4 is located downstream from the mouth of the Main Drainage Channel of Zagreb, where the impact of discharged Zagreb wastewaters was significant. Twelve Trans National Monitoring Network (TNMN) stations are operating in the Sava River Basin, among them three stations are on the Sava River in Croatia (S1, S8, and S17).

Sampling and chemical analysis

Water quality of the Sava River is monitored regularly in accordance with national regulations. In this study, 18 physico-chemical parameters obtained at each station were used for analysis. Data for the sampling period from 01 January 2000 to 31 December 2006 are presented when samples were collected twice per

Fig. 1 Map of the sampling site locations



month (every other week). Sampling, preservation, and chemical analyzes were performed according to standard analytical methods for the examination of surface waters (ISO; APHA 1995; EPA 1999) which are routinely applied in the water quality monitoring laboratories. Air and water temperatures were measured on the site while pH, electrical conductivity, and dissolved oxygen were measured both on the site (by using standard electrochemical techniques; portable pH, dissolved oxygen and conductivity meters, HANNA instruments) and in the laboratory. For meaningful data interpretation, water quality parameters measured at various temperatures need to be transformed to values corresponding to a standard temperature (Hayashi 2004). Therefore, only data measured in the laboratory were used here. Unless otherwise specified in the methods, water samples were collected in 3–5 L polyethylene containers, stored in the dark at 4 °C, and analyzed within 24 h. For the total suspended solids determination, samples were filtered through 0.45 cellulose nitrate membrane filter (Whatman, Springfield Mill, England). The water quality parameters, their units and abbreviation and methods of analysis are summarized in Table 1.

Although at most stations, samples have been collected over 30 years and there are more than 50 water quality parameters available (including detergents, pesticides, and other organic parameters, biological and microbiological parameters, heavy metals, etc.), only 26 parameters have been selected due to the continuity in measurement at all water quality monitoring stations. Constituents that were not routinely analyzed and/or analyzed at all stations were excluded. This data set was further refined by retaining only 18 parameters; highly redundant constituents (such as multiple forms of nitrogen) and constituents that have nearly constant concentrations over the entire watershed were excluded. Finally, we have 18 parameters at 18 locations over the 7-year period.

Data and statistical methods

Data set

The data set used in this work consisted of 3164 samples by 18 variables. This rather large set of data (a total of 56,952 values) contained only eight missing data (0.01 %) which were replaced with median values

Table 1 Determined water quality parameters and analytical methods used in the study

VARIABLE	ABBREVIATION	UNITS	METHODS
Air temperature	A-TEMP	°C	
Water temperature	W-TEMP	°C	SM 2550
pH	pH		ISO 10523:1998
Electrical conductivity	EC	μS/cm	ISO 7888:1985
Total suspended solids	TSS	mg/L	ISO 11923:1997
Alkalinity	ALK	mgCaCO ₃ /L	ISO 9963-1:1994
Dissolved oxygen	DO	mgO ₂ /L	ISO 5813:1983
Oxygen saturation	OS	%	
Permanganate index	COD _{Mn}	mgO ₂ /L	ISO 8467:1993
Chemical oxygen demand	COD _{Cr}	mgO ₂ /L	ISO 6060:1989
5-d biological oxygen demand	BOD ₅	mgO ₂ /L	ISO 5815:1998
Ammonium	NH ₄ ⁺	mgN/L	ISO 7150-1:1984
Nitrite	NO ₂ ⁻	mgN/L	SM 4500-NO ₂
Nitrate	NO ₃ ⁻	mgN/L	SM 4500-NO ₃
Kjeldahl nitrogen	TKN	mgN/L	SM 4500-Norg
Total phosphorus	TP	mgP/L	SM 4500-P
Total oil	TO	mg/L	EPA 1664*
Mineral oil	MO	mg/L	EPA 1664*

* modified

of appropriate variables (Olsen et al. 2012) and 234 data (0.4 %) which were found to be below detection limit and were replaced with half of the detection limit value (Farnham et al. 2002). Basic statistical properties of the data set used are summarized in Table 2. Along with the distribution of missing data (MD) and the data below the detection limit (DL), the moments of the data distributions are given. The values of skewness and kurtosis of all variables but A-TEMP, W-TEMP, EC, and DO depart significantly from the normal distribution. Hence, to describe the distribution of the data in this work, the robust statistic, median, and interquartile range (IQR) are used (right hand-side of the Table 2). All data whose values lay outside the interval ($Q_1 - 3 \cdot IQR, Q_3 + 3 \cdot IQR$), where Q_1 and Q_3 are the first and third quartile, respectively, were considered as outlying. According to this criterion, there were 1.7 % univariate and 3.8 % multivariate outliers in the data set. The number of multivariate outliers was determined by applying described criterion on the distribution of Mahalanobis distances from common center (Mahalanobis 1936). We do not want to suggest that those values are indicative of measurement errors or bias in the data but to emphasize that

populations have skewed and heavy-tailed distributions and that tools that assume a normal distribution should be used with caution.

Principal components analysis (PCA)

PCA is a mathematical procedure that uses orthogonal transformation to convert a set of observations into a set of linearly uncorrelated variables called principal components. Due to the quadratic error criterion, standard PCA algorithm is very sensitive and its output can change dramatically in the presence of only a few outliers (Maronna et al. 2006). The data set used in this work contained variables whose distributions strongly deviate from normal distribution, with some having strongly skewed and heavy tailed distributions and/or non negligible number of outliers (see Table 2). Therefore, in addition to standard algorithm, the robust PCA algorithm was used. The PCA was performed on covariance matrix that was calculated by weighting the mean and the products which form it. The data were weighted with respect to their distance from the center. Two ways of weighting were used: univariate where the variables were

Table 2 Descriptive statistics of the water quality parameters

Variable	MD ^a	DL ^b	Mean	Variance	Skewness	Kurtosis	Median	IQR ^c	OP ^d
A-TEMP	4		14.3	92.6	-0.231	2.46	15.0	14.5	0
W-TEMP			14.2	50.8	0.160	1.97	13.6	12.0	0
pH			7.89	0.077	-0.694	4.04	7.90	0.320	0
EC			407	3005	0.243	3.07	402	75.0	0
TSS		1	21.9	2121	11.2	196	12.0	14.0	5.31
ALK	1		213	2837	2.35	8.80	200	33.0	7.93
DO	1		8.75	3.68	-0.191	3.70	8.80	2.75	0
OS			83.8	287	2.04	16.7	82.9	13.3	2.75
COD _{Mn}			4.22	4.17	9.21	193	3.80	1.54	1.30
COD _{Cr}	1		13.3	41.9	3.26	31.8	12.3	6.20	0.88
BOD ₅			2.73	1.69	1.80	9.36	2.50	1.40	0.88
NH ₄ ⁺		4	0.175	0.034	3.01	21.1	0.120	0.170	1.52
NO ₂ ⁻			0.030	0.0006	7.27	121	0.024	0.018	1.96
NO ₃ ⁻			1.34	0.253	1.08	8.41	1.30	0.600	0.35
TKN			0.738	0.319	2.06	11.4	0.590	0.650	0.54
TP	1		0.193	0.020	3.15	19.6	0.160	0.120	2.05
TO		14	0.170	0.047	6.43	69.6	0.130	0.149	2.53
MO		215	0.052	0.009	21.8	719	0.040	0.041	1.68

^anumber of missing data, ^bnumber of data below the detection limit, ^cinterquartile range, ^doutliers percentage (univariate)

weighted independently and multivariate where all variables were weighted simultaneously. In the first case, the weight of each data point was calculated as inverse value of absolute distance from median in the units of median of absolute distances for a given variable:

$$w_i = \frac{\text{median}(|x_i - \text{median}(x_i)|)}{|x_i - \text{median}(x_i)|} \quad (1)$$

where w_i is the weight, x_i are measurements of given variable, and $||$ denotes absolute value. In this way, matrix of weights w_{ij} of the same dimension as the data matrix was obtained. In the multivariate case, the data were weighted with the normalized reciprocal value of Mahalanobis distance. In this way, obtained vector w_i obtained in this way was used in calculation of covariance matrix.

Cluster analysis (CA)

CA is a common technique for statistical data analysis and exploratory data mining used in many fields as well in water quality assessment (Shrestha and Kazama 2007; Kazi et al. 2009). With the aim of examining spatial variability (grouping similar sampling sites and spreading them over the river), in this study, hierarchical agglomerative clustering was performed on the normalized data set by the Ward's method of linkage with squared Euclidean distances as a measure of similarity. Unlike PCA which uses a number of principal components for display purposes, CA uses all the information contained in the original data set.

Results and discussion

Correlation

The correlation matrix of the 18 variables was calculated. The obtained correlation coefficients should be interpreted with caution because they reflect both spatial and temporal variations (see Fig. 2). Strong correlations between A-TEMP and W-TEMP ($R = 0.89$), COD_{Mn} and COD_{Cr} ($R = 0.71$), TSS and both CODs ($R = 0.70$, $R = 0.54$), as well as DO and OS ($R=0.69$) and DO and temperatures ($R=-0.60$, $R = -0.51$) can be observed. Somewhat weaker correlations are observed between both CODs and BOD_5 , TO and MO, NH_4^+ and TKN, and NO_3^- and temperatures. The obtained

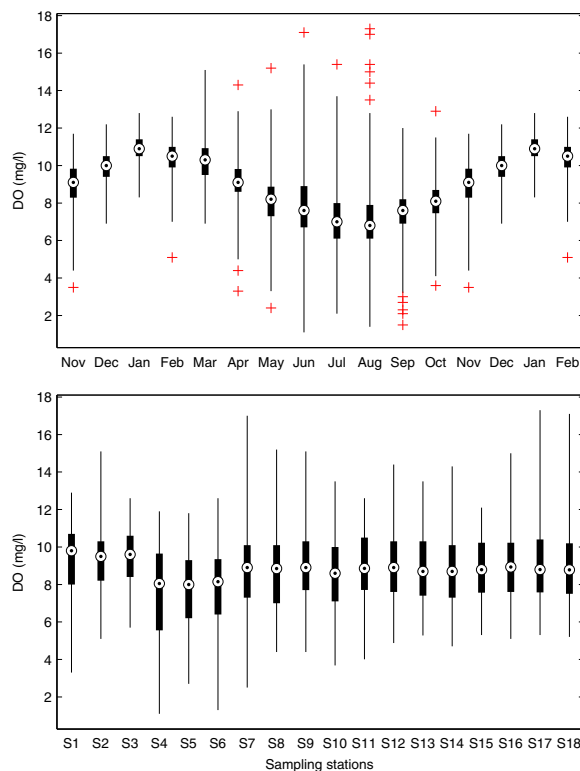


Fig. 2 Temporal and spatial variations of DO

results are expected and consistent with the fact that these variables are dependent on each other.

The relationship between climatic, hydrological, and water quality parameters was recently studied by other authors. Most of them analyzed influence of air temperature on river water quality. Our results match well with the results of those studies. Analyzing meteorological and river water quality data Ozaki et al. (2003) found that an increase in A-TEMP resulted in rises in BOD and SS and a drop in DO. From the analyzes of the four seasons, the gradients differed significantly over the seasons for W-TEMP, BOD, DO, DO saturation ratio, SS and pH, and enhanced dependencies were observed in summer. In our work, enhanced A-TEMP dependencies of W-TEMP and BOD are observed in spring and DO and OS in autumn. It should be mentioned that Ozaki et al. (2003) used slightly different boundaries for seasons than the ones used here. Studying the lower Mekong River Prathumratana et al. (2008) found positive correlation between A-TEMP and TSS, NO_3^- , TP and COD and negative correlations with DO and ALK. Similar results were

obtained in this paper except for the negative correlation between A-TEMP and NO_3^- . However, studying climate-water quality relationships in large rivers on a global scale by investigating the climate elasticity of river water quality Jiang et al. (2014) concluded that some parameters can be site specific (in their work NO_2^- and ortho-P). They also showed that elasticity is basically consistent with the results of statistical correlation on the direction of response. Correlations obtained here are in the same direction as the elasticity obtained by Jiang et al. (2014).

Multivariate analysis of the entire data set

Prior to PCA, data were normalized to have zero means and unit variances to account for different measurement units and to equalize the impact of all variables on the total variance in the data set. Seven principal components having eigenvalues larger than one were obtained explaining over 75 % of the total variance (with individual components explaining 20.0, 13.4, 13.1, 8.8, 7.7, 6.4, and 6.2 %). To simplify the structure of the obtained solution, varimax rotation of the seven component subspace was performed and the resulting variable loadings are given in Table 3. Simultaneously with simplifying the loadings

structure of the components, varimax rotation equalizes their variances making previously more important components less important and vice versa. The percent of total variance explained by first and seventh component changed from 20.0 and 6.2 % to 14.3 and 8.3 %, respectively.

The first rotated component (VF1) is responsible for 14.3 % of the total variance and has strong (Liu et al. 2003) negative loadings for air and water temperatures and moderate positive loadings for DO and NO_3^- suggesting a physico-chemical source of variability. This varifactor can be explained that temperature increase leads to decrease the amount of DO and acceleration of nitrification in water. It also indicates that the most of the variability in the data is due to the temperature changes. VF2 accounts for 14.2 % of the total variance and is correlated primarily with TSS, COD_{Mn} , and COD_{Cr} . This component represents influences from organic pollutants related mostly to human activities, such as domestic and industrial discharges, and also to decaying plant and animal matter. VF3 accounting for 10.9 % of the total variance is largely contributed by DO and OS and represents physico-chemical source of the variability and influences from natural inputs. The total explained variance of VF4 is 10.8 %. It has a strong loading

Table 3 Loadings of 18 variables on Varimax rotated principal components

	1	2	3	4	5	6	7
A-TEMP	-0.924	-0.018	-0.046	0.014	-0.067	0.053	-0.040
W-TEMP	-0.937	-0.001	-0.044	0.111	-0.046	0.042	-0.018
pH	-0.129	0.164	0.434	0.014	0.224	0.568	0.138
EC	-0.052	0.412	0.121	0.768	-0.056	0.202	-0.001
TSS	0.069	-0.921	-0.018	-0.229	-0.031	-0.056	0.039
ALK	0.012	0.085	0.167	0.215	0.057	-0.859	0.048
DO	0.549	-0.009	0.773	-0.018	-0.060	-0.015	-0.022
OS	-0.162	-0.030	0.917	0.067	-0.113	0.004	-0.042
COD_{Mn}	-0.060	-0.884	-0.018	0.162	-0.089	-0.056	0.011
COD_{Cr}	-0.010	-0.709	0.040	0.171	0.102	0.356	-0.022
BOD_5	-0.077	-0.384	0.196	0.456	0.166	-0.046	-0.099
NH_4^+	0.287	0.091	-0.229	0.264	0.675	-0.020	0.025
NO_2^-	-0.068	0.009	-0.208	0.693	-0.105	-0.102	0.033
NO_3^-	0.610	-0.035	-0.368	0.254	-0.235	0.283	-0.141
TKN	-0.113	-0.030	0.029	-0.146	0.916	0.074	-0.055
TP	0.052	-0.241	0.070	0.572	0.118	-0.115	0.039
TO	-0.077	-0.015	0.086	-0.111	0.165	-0.333	-0.759
MO	0.029	0.052	-0.035	0.064	-0.110	0.248	-0.926

of EC and moderate of inorganic nutrients (NO_2^- and TP). This inorganic nutrient-type component could be interpreted as representing the influences from agricultural chemical application (excessive use of fertilizers), domestic and industrial discharges, and the erosion of natural deposits. VF5 (8.6 % of the variance) that is weighted on water-soluble nitrogen species (TKN and NH_4^+) is likely to represent the sewage and manure discharges to water bodies, influences from agricultural runoff and byproducts from the industrial manufacturing processes. VF6 (8.5 % of the variance) is negatively loaded on ALK and positively on pH, representing physico-chemical source of the variability (mineral component of the water). VF7 explaining 8.3 % of the total variance is correlated with oils. This component could be interpreted as representing influences from municipal and industrial sewage, runoff from roads and municipal areas. Also, ships and motorboats might contribute significantly to water pollution with oils.

These results are consistent with the results of other authors which have used PCA for water quality investigations. In order to characterize the nature of the water quality impairment in the Wen-Rui Tang River watershed (China) and the relationships among the water quality parameters, Mei et al. (2014) selected the ten correlated parameters for FA. FA identified three factors with eigenvalues >0.96 (1) summing to 70.9 % of the total variance in the water quality dataset. Results revealed that parameters related to organic pollutants (VF1), nutrients (VF2), and salt concentration (VF3) were the most important parameters contributing to water quality variation. Pinto and Maheshwari (2011) employ FA to considerably reduce the number of variables obtained in a routine monitoring program in the Hawkesbury-Nepean River system in New South Wales, Australia and identify the latent factors relative to river health in peri-urban landscapes. Out of 40 water quality variables measured on monthly basis during 2008 and 2009 included in the analysis, the FA identified nine key variables, under three varifactors (VFs), explaining 50 % of the variance in the river water quality. Anaerobic fermentation, microbial pollution, and eutrophication are three key environmental problems faced by peri-urban rivers. In an attempt to differentiate between sources of variation in the water quality of the Ebro River (Spain), Bouza-Deano et al. (2008) carried out exploratory analysis of data by CA and

PCA. In their study, PCA has allowed the identification of the following factors: geologic (VF1 and VF2), climatic (VF3), and anthropogenic (VF4). Simeonov et al. (2003) employed PCA for the interpretation of a large and complex data matrix (21 parameters determined at 25 sampling sites for a period of 36 months) obtained during a monitoring program of surface waters in Northern Greece. Six principal components were obtained. VF1, VF2, and VF3 represent organic, nutrient, and physico-chemical source of the variability and are similar to components obtained in our study. Similar results were obtained by other authors which were studying surface water quality using multivariate techniques (Ouyang et al. 2006; Fan et al. 2010; Razmkhah et al. 2010; Nasir et al. 2011).

In the Fig. 3, the VF1/VF2 scores plot is given. Individual scores (a) as well as sampling stations (b) and seasonal (c) medians are shown. The variances of the VFs are denoted with the white ellipse in the panel (a) whose semimajor axes are equal to standard deviations of VFs. It can be noted that all of the points are located roughly within 3σ in the VF1 direction while this is not the case in the VF2 direction. This reflects the structure of those components: VF1 is composed from variables whose distributions are close to normal distribution, while VF2 contains variables whose distributions differ significantly from the normal one. More negative values of VF1 indicate higher temperature, while more negative values of VF2 represent higher organic pollution. It can be readily seen that S4 has highest organic pollution. As already mentioned, station S4 is located downstream from Zagreb, the largest municipal center in the watershed. The temperatures and the organic pollution are higher in summer and spring than in autumn and winter, as expected. It can be noted on the figure that upper stations (S1–S4) have higher values of VF1 than the downstream stations indicating lower temperatures at those stations. This can be explained by the fact that samples from the stations near Zagreb have been collected in the morning while the stations downstream have been sampled mostly in the afternoon when the air temperature is higher. Stations S1–S4 have higher NO_3^- values and stations S1, S2, and S3 additionally higher DO median what also leads to increased VF1 median for these stations.

On the banks of the Sava River and its tributaries are located cities with developed industries

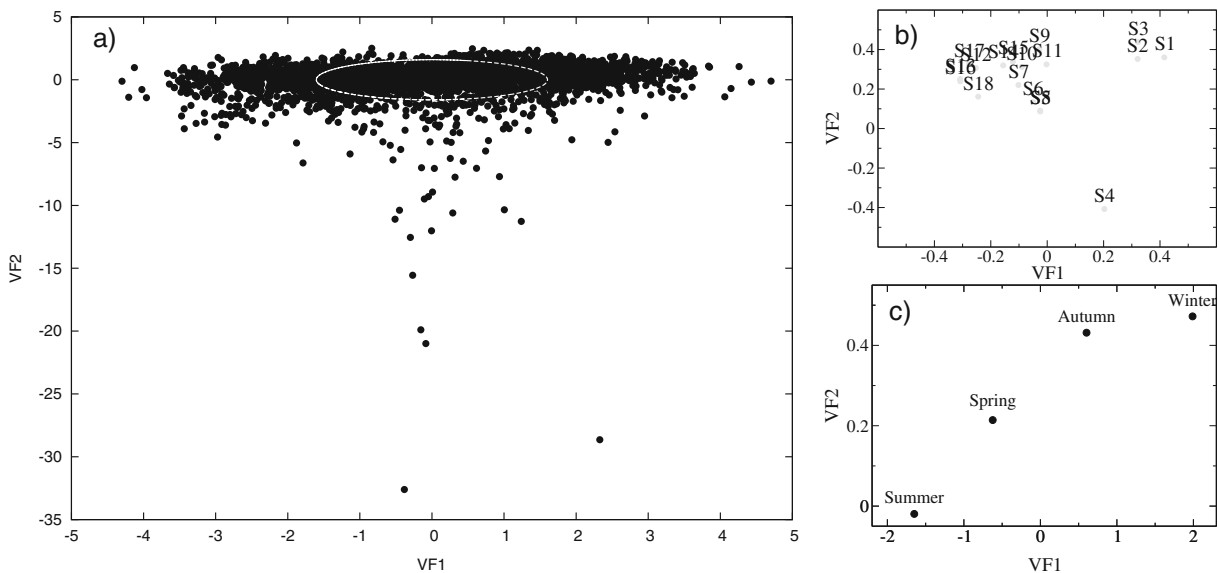


Fig. 3 VF1 versus VF2 scores for the Sava River: (a) individual scores, (b) sampling stations, and (c) seasonal medians

such as Zagreb - capital city of Croatia with significant population and diverse industry, Krško - nuclear power plant, Sisak - river port, refinery, food factory, factory of alcoholic beverages, Kutina - petrochemistry, Slavonski Brod - food processing, metal processing, Bosanski Brod - oil refinery, Županja - sugar factory, wood industry, food industry, etc. Also, agriculture and animal breeding followed by food industry are well developed especially in the middle and lower part of the Sava watercourse. The main sources of organic and nutrient pollution come from untreated wastewaters discharged from municipalities and industries along the Sava River as well as the tributaries. For the year 2007, in Croatia, 104 agglomerations ≥ 2000 PE in the Sava River Basin were present out of which 89 agglomerations emitted wastewaters into the environment without any treatment.

Examination of spatial changes of rotated components reveals that all VF, except VF1, have large extreme values at station S4 which in most cases approach median through station S5 and S6. Therefore, we can conclude that the only notable (significantly above the error) source of pollution is Main Drainage Channel of Zagreb and its untreated wastewaters. All other changes of VF are of the order or below the error level so the impact of corresponding sources of pollution cannot be undoubtedly identified in our results.

To further analyze spatial variations of water quality and the similarity of the sampling sites, data were subjected to CA. The dendrogram obtained by Ward's method is shown in Fig. 4. The resulted dendrogram grouped all 18 sampling sites into four statistically significant clusters at $(D_{link}/D_{max}) \cdot 100 < 70\%$. As in the work Wang et al. (2012), Li et al. (2011), Bouza-Deano et al. (2008), and Shrestha and Kazama (2007), the clustering procedure highlighted groups in which the sites have similar characteristics and natural source types. The cluster I (stations S10–S18) is situated at

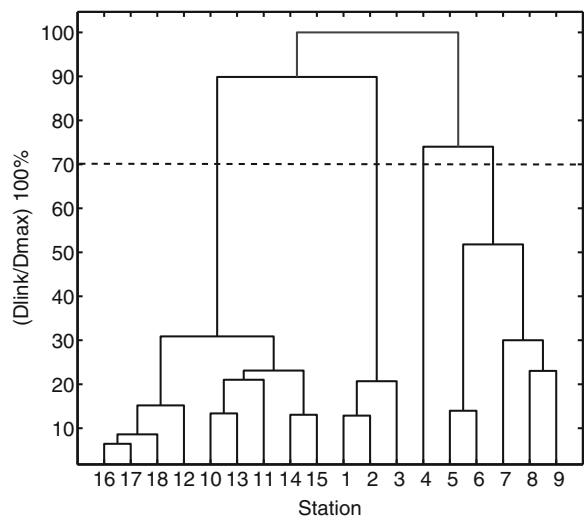


Fig. 4 Dendrogram obtained by cluster analysis

the most downstream site of the river and corresponds to relatively low polluted regions. On the other hand, cluster II is situated at the most upstream site of the river; it covers an area upstream of Zagreb (S1 and S2) and Zagreb (S3) and corresponds to moderately polluted region. Cluster III, i.e., station S4 where the impact of discharged Zagreb wastewaters were significant, corresponds to the lowest water quality site. Cluster III is associated with cluster IV (sites S5–S9) suggesting the impact of pollution on downstream stations. The cluster IV corresponds to moderately polluted sites and represents stations which are influenced by upstream pollution and also by the pollution of the city of Sisak.

The results obtained in this section can be used to reduce the number of variables and/or stations in order to reduce the number of analyzes and the costs. For rapid quality assessment studies, number of the sampling sites could be reduced and only representative sites from each cluster identified by CA could be used. The same was already suggested by, e.g., Simeonov et al. (2003) and Bouza-Deano et al. (2008).

The PCA combines variables of 'similar' patterns into groups indicating that those variables are dependent on each other or may have the same background. For each group of variables, the key or principal variable can be chosen as an representative. In this way, the number of variables could be reduced. In our case, we can identify pairs: COD_{Mn} and COD_{Cr} , TO and MO, and DO and OS. One member could be used as representative for each pair, so we could try to omit COD_{Cr} , TO, and OS. However, one has to be very careful when applying this procedure because the variables with different causes may be combined in a single group as pointed out by Weilguni and Humpesch (1999).

Multivariate analysis of the seasonal subsets

To further examine the influence of seasonal changes of water quality parameters, the above described analysis was repeated for each season separately. Similar analysis was performed by Wang et al. (2012), Li et al. (2011), Ouyang et al. (2006), and Razmkhah et al. (2010). Ouyang et al. (2006) who investigated the seasonal variations in water quality of the lower St. Johns River in Florida, USA using the PCA and PFA techniques found that water quality parameter that is most important in contributing to water quality

variation for one season may not be important for another.

The data set was divided into four subsets (spring, summer, autumn, and winter) which were again normalized to have zero mean and unit variance. PCA was performed and, most probably due to smaller data set size, only six PCs had eigenvalues greater than one explaining 73.6 % (spring), 71.4 % (summer), 70.3 % (autumn), and 71.3 % (winter) of the total variance.

The results for individual seasons are similar to those obtained for the whole data set. The most important parameters of varimax rotated solutions in all seasons are TSS and CODs (see Table 4) in contrast to the temperatures which were the most important parameters for the whole data set. The decrease of contribution of physical factor to the total variance in season subsets is expected as the temperature changes are significantly smaller within seasons (during spring or summer) than between the seasons (summer and winter). On the other hand, the importance of other parameters (TKN, NH_4^+ , and TP) has increased.

Furthermore, the results for seasonal subsets are qualitatively similar with the results recently obtained by Wang et al. (2012) who used PCA to explore the most important factors determining the spatiotemporal dynamics of water quality in Xiangxi River. They analysed a 5-year (2002–2006) continual monitoring data (14 parameters at 12 sites). PCA of the four data sets yielded six PCs for spring and autumn and five PCs for summer and winter with eigenvalues >1 , explaining 74.70, 74.47, 71.02, and 69.86% of the total variance in respective water quality data sets. The VFs obtained from the PCs suggested that the parameters responsible for water quality variations are mainly related to the dilution of salt (natural), the point source pollution of phosphorus and the non-point pollution of nitrogen (anthropogenic).

To investigate water quality in the rivers along the water conveyance canal of the Middle Route of China's interbasin South to North Water Transfer Project and to assess the spatial and temporal patterns of water quality in the rivers Li et al. (2011) used multivariate statistical analyses. PCA extracted four, five, five, and three principal components (PCs) for the measurement data in September, December, April, and June, respectively, and these PCs with eigenvalue >1 explained 79, 83, 88, and 83% of the total variance in the respective data sets. Seasonal FA/PCA allowed

Table 4 Rotated factor correlations with water quality parameters for whole dataset, univariate and multivariate weighted data and for seasonal subsets

Data	Correlation	VF1	VF2	VF3	VF4	VF5	VF6	VF7
All	High	TEMP	TSS,COD _{Mn}	DO,OS	EC	TKN	ALK	MO,TO
	Moderate	DO,NO ₃ ⁻	COD _{Cr}		NO ₂ ⁻ ,TP	NH ₄ ⁺	pH	
Univ.	High	COD _{Mn} ,COD _{Cr} , BOD ₅	TEMP	DO,OS	ALK	TKN,NH ₄ ⁺	EC,TSS	TO,MO
	Moderate	NO ₂ ⁻	DO,NO ₃ ⁻	pH,NO ₂ ⁻ ,NO ₃ ⁻	COD _{Cr} ,NO ₃ ⁻	pH,TO	NO ₂ ⁻ ,pH,NO ₃ ⁻	pH
Multiv.	High	COD _{Mn} ,COD _{Cr} , BOD ₅	TEMP	DO,OS	EC	ALK	TKN,NH ₄ ⁺	TO,MO
	Moderate	TSS,TP	DO,NO ₃ ⁻		TSS,NO ₂ ⁻	pH,COD _{Cr}		pH
Spring	High	TSS,COD _{Mn}	DO,OS	TEMP	MO	ALK		
	Moderate	COD _{Cr} ,BOD ₅ ,		NO ₃ ⁻	TO,NH ₄ ⁺	TO,pH	EC,TKN,NO ₂ ⁻	
Summer	High	TSS,COD _{Mn} , COD _{Cr}	EC	DO,OS	NO ₂ ⁻		MO	
	Moderate		TEMP	NH ₄ ⁺	pH,ALK, BOD ₅ ,TP	TKN,NO ₃ ⁻	TO	
Autumn	High	COD _{Cr} ,COD _{Mn} , TSS	TEMP	TKN,NH ₄ ⁺	EC	TO,MO	pH	
	Moderate	BOD ₅	DO	OS	NO ₂ ⁻		ALK	
Winter	High	TSS,COD _{Mn}	OS,NH ₄ ⁺ ,TKN	TEMP		TO,MO	ALK	
	Moderate	COD _{Cr} ,TP	DO	DO	EC,NO ₂ ⁻ ,NO ₃ ⁻		NO ₃ ⁻	

four categories of parameters such as mineral composition (primarily natural), toxic metals (industrial), nutrients (agricultural, domestic, and industrial), and organic pollutants (domestic, municipal, and industrial sources).

Robust PCA

To investigate the influence of the skewed, heavy tailed data distributions containing outliers on the results, the robust PCA algorithm was applied. This was performed by calculating the robust estimates of covariance matrix which were then subjected to PCA. Since covariance and correlation differ by the factor $1/(\sigma_x \cdot \sigma_y)$, and the data set was normalized to have unit variances (and therefore standard deviations), the calculated covariances can be compared to the correlations obtained earlier and examined for the effects of weighting. The largest differences are in correlations between TSS and both CODs which decreased for about 0.2 in all cases. The decrease is more pronounced in the case of univariate weighting with maximum near 0.3 for the correlation between TSS

and COD_{Mn}. Similar but somewhat smaller change of correlation (about 0.1 decrease) is observed for correlation between COD_{Mn} and COD_{Cr}. The correlation between DO and temperatures and correlation between NH₄⁺ and TKN increased for about 0.1 with the increase more pronounced in the multivariate case for the former and in the univariate case for the later correlation. Other significant correlations remain unchanged.

The PCA yielded seven principal components explaining 75 % of the total variance, with individual components explaining 21.6, 14.1, 11.6, 8.3, 7.5, 6.8, and 5.7 % (univariate weighting) and 20.4, 14.6, 11.5, 8.1, 7.5, 6.7, and 6.2% (multivariate weighting). These results are the same as the result obtained by standard PCA algorithm. Again, the varimax rotation of seven component subspace was performed. Correlations between obtained rotated varifactors and water quality parameters are given in Table 4. Figure 5 shows the PC1 versus PC2 variable loading vectors (prior to varimax rotation) for data without weights (a), univariate (b), and multivariate (c) weighted data and for the spring data subset (d). The results of

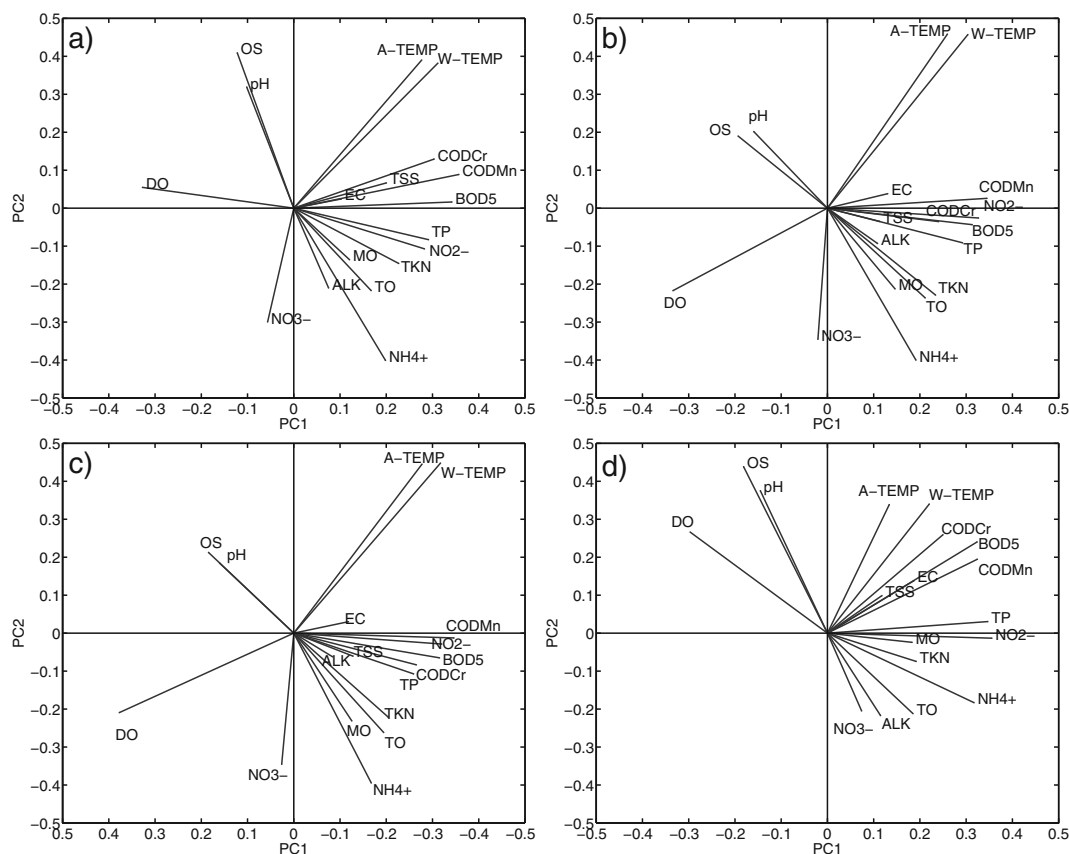


Fig. 5 PC1 versus PC2 loadings for the Sava River (see Table 1 for acronym identification); a) data without weights, b) univariate and c) multivariate weighted data, d) spring data subset

standard and robust PCA are fairly similar. Actually, the differences between the standard and robust PCA algorithm (univariate or multivariate weighting) are smaller than the differences between any of them and spring subset.

Conclusion

In this study, surface water quality data for 18 parameters collected from 18 monitoring stations along the Sava River in Croatia from 2000 to 2006 were analyzed using multivariate statistical techniques.

Most of the variables were found to have skewed heavy tailed distributions containing outliers. It should be emphasized that the standard statistical methods assume normal distribution and should be used with caution when analysing the data whose distributions are significantly departing from normal distribution.

When analysing the correlations between variables, correlations between air temperature and some water quality parameters were found in agreement with the previous studies of relationship between climatic and hydrological parameters. Further significant correlations between TSS and CODs were found but they are diminishing when the weighting procedure is applied.

PCA has determined a reduced number of seven principal components that explain over 75 % of the data set variance. Varimax rotation of the seven components subspace resulted in simpler structure of rotated components each of them related to small group of measured water quality parameters.

Cluster analysis has found similarities in sampling stations across the river. All 18 sampling sites have been grouped into four statistically significant clusters. The first group is located at the end and the second group at the top of the river. These two groups

correspond to relatively low and moderately polluted sites. The third and fourth group which correspond to highly and moderately polluted sites are located in the middle of the river.

The PCA analysis of the seasonal subsets showed that the importance of the parameters is changing from season to season and that the parameters which are contributing most to the water quality variation in one season could be contributing less (or not at all) in another season. However, no significantly different results were found for any season when compared with the results for the whole data set. Changes of order or sign in some components like the increased importance of organic component and the decrease of the physical component in all seasons does not represent significantly different result. Therefore, these results give rise to the reliability of the obtained results.

Those results are similar to those obtained by other authors who applied statistical analysis techniques such as Pearson’s correlation, PCA, and CA for the analysis of the data obtained in the water quality monitoring programs.

The temperatures and CODs are found to be the parameters which are responsible for most of the variance in the data set indicating that the physical- and organic-related sources contribute mostly to the water quality variations.

Finally, to check the influence of the outliers in the data set whose distribution strongly deviates from the normal one, two robust estimates of covariance matrix were calculated and subjected to PCA. Again, no significant differences between the obtained results were found. Hence, in the case of the data set with small number of missing data, non-detect values and outliers (less than 4 %), the usage of standard PCA algorithm is justified.

Acknowledgments We wish to thank the anonymous referees whose detailed and very careful review of the manuscript helped us to significantly improve the quality of this paper.

Conflict of interests The authors declare that they have no conflict of interest.

References

APHA. (1995). *Standard Methods for the Examination of Water and Waste Water*, 19th edn. Washington, DC: American

Public Health Association, American Water Works Association and Water Pollution Control Federation.
 Bouza-Deano, R., Ternero-Rodriguez, M., & Fernandez-Espinosa, A.J. (2008). Trend study and assessment of surface water quality in the Ebro River (Spain). *Journal of Hydrology*, 361(3–4), 227–239. doi:10.1016/j.jhydrol.2008.07.048.
 Campbell, N. A. (1980). Robust procedures in multivariate-analysis. I. Robust covariance estimation. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 29(3), 231237. doi:10.2307/2346896.
 Croux, C., & Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3), 603–618. doi:10.1093/biomet/87.3.603.
 Dixon, W., & Chiswell, B. (1996). Review of aquatic monitoring program design. *Water Research*, 30(9), 1935–1948. doi:10.1016/0043-1354(96)00087-5.
 EPA. (1999). *Method 1664, EPA-821-R-98-002*. Washington, DC: United States Environmental Protection Agency.
 Fan, X., Cui, B., Zhao, H., Zhang, Z., & Zhang, H. (2010). Assessment of river water quality in Pearl River Delta using multivariate statistical techniques. *Procedia Environmental Sciences*, 2, 1220–1234. doi:10.1016/j.proenv.2010.10.133.
 Farnham, I., Singh, A., Stetzenbach, K., & Johannesson, K. (2002). Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems*, 60(1–2), 265–281. doi:10.1016/S0169-7439(01)00201-5. 4th International Conference on Environmetrics and Chemometrics, LAS VEGAS, NEVADA, SEP 08–20, 2000.
 Hayashi, M. (2004). Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data inversion. *Environmental Monitoring and Assessment*, 96(1–3), 119–128. doi:10.1023/B:EMAS.0000031719.83065.68.
 Hubert, M., Rousseeuw, P., & Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1–2), 101–111. doi:10.1016/S0169-7439(01)00188-5. 4th International Conference on Environmetrics and Chemometrics, LAS VEGAS, NEVADA, SEP 08–20, 2000.
 Jiang, J., Sharma, A., Sivakumar, B., & Wang, P. (2014). A global assessment of climate-water quality relationships in large rivers: an elasticity perspective. *Science of the Total Environment*, 468, 877–891. doi:10.1016/j.scitotenv.2013.09.002.
 Kazi, T.G., Arain, M.B., Jamali, M.K., Jalbani, N., Afridi, H.I., Sarfraz, R.A., Baig, J.A., & Shah, A.Q. (2009). Assessment of water quality of polluted lake using multivariate statistical techniques: a case study. *Ecotoxicology and Environmental Safety*, 72(2), 301–309. doi:10.1016/j.ecoenv.2008.02.024.
 Li, S., Li, J., & Zhang, Q. (2011). Water quality assessment in the rivers along the water conveyance system of the Middle Route of the South to North Water Transfer Project (China) using multivariate statistical techniques and receptor modeling. *Journal of Hazardous Materials*, 195, 306–317. doi:10.1016/j.jhazmat.2011.08.043.

- Liu, C.W., Lin, K.H., & Kuo, Y.M. (2003). Application of factor analysis in the assessment of ground-water quality in a blackfoot disease area in Taiwan. *Science of the Total Environment*, 313(1–3), 77–89. doi:10.1016/S0048-9697(02)00683-6.
- Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1), 49–55.
- Malmqvist, B., & Rundle, S. (2002). Threats to the running water ecosystems of the world. *Environmental Conservation*, 2, 134–153. doi:10.1017/S0376892902000097.
- Maronna, R., Martin, R.D., & Yohai, V.J. (2006). *Robust statistics: Theory and methods*. Chichester: John Wiley & Sons.
- Mayer, B., Boyer, E.W., Goodale, C., Jaworski, N.A., Van Breemen, N., Howarth, R.W., Seitzinger, S., Billen, G., Lajtha, L., Nosal, M., & Paustian, K. (2002). Sources of nitrate in rivers draining sixteen watersheds in the north-eastern US: Isotopic constraints. *Biogeochemistry*, 57(1), 171–197. doi:10.1023/A:1015744002496.
- Mei, K., Liao, L., Zhu, Y., Lu, P., Wang, Z., Dahlgren, R.A., & Zhang, M. (2014). Evaluation of spatial-temporal variations and trends in surface water quality across a rural-suburban-urban interface. *Environmental Science and Pollution Research*, 21(13), 8036–8051. doi:10.1007/s11356-014-2716-z.
- Meybeck, M. (1998). Man and river interface: multiple impacts on water and particulates chemistry illustrated in the Seine river basin. *Hydrobiologia*, 374, 1–20. 3rd International Joint Conference on Limnology and Oceanography: Oceans, Rivers and Lakes - Energy and Substance Transfers at Interfaces, NANTES, FRANCE, OCT, 1996.
- Nasir, M.F.M., Samsudin, M.S., Mohamad, I., Awaluddin, M.R.A., Mansor M.A., Juahir, H., & Ramli, N. (2011). River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): A case study at Klang River, Malaysia. *World Applied Sciences Journal*, 14, 73–82.
- Olsen, R.L., Chappell, R.W., & Loftis, J.C. (2012). Water quality sample collection, data treatment and results presentation for principal components analysis - literature review and Illinois River watershed case study. *Water Research*, 46(9), 3110–3122. doi:10.1016/j.watres.2012.03.028.
- Ouyang, Y., Nkedi-Kizza, P., Wu, Q.T., Shinde, D., & Huang, C.H. (2006). Assessment of seasonal variations in surface water quality. *Water Research*, 40(20), 3800–3810. doi:10.1016/j.watres.2006.08.030.
- Ozaki, N., Fukushima, T., Harasawa, H., Kojiri, T., Kawashima, K., & Ono, M. (2003). Statistical analyses on the effects of air temperature fluctuations on river water qualities. *Hydrological Processes*, 17(14). doi:10.1002/hyp.1437.
- Pinto, U., & Maheshwari, B.L. (2011). River health assessment in pen-urban landscapes: an application of multivariate analysis to identify the key variables. *Water Research*, 45(13), 3915–3924. doi:10.1016/j.watres.2011.04.044.
- Prathumratana, L., Sthiannopkao, S., & Kim, K.W. (2008). The relationship of climatic and hydrological parameters to surface water quality in the lower Mekong River. *Environment International*, 34(6), 860–866. doi:10.1016/j.envint.2007.10.011.
- Razmkhah, H., Abrishamchi, A., & Torkian, A. (2010). Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: a case study on Jajrood River (Tehran, Iran). *Journal of Environmental Management*, 91(4), 852–860. doi:10.1016/j.jenvman.2009.11.001.
- Ruymgaart, F.H. (1981). A robust principal component analysis. *Journal of Multivariate Analysis*, 11(4), 485–497. doi:10.1016/0047-259X(81)90091-9.
- Shrestha, S., & Kazama, F. (2007). *Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan* (Vol. 22, pp. 464–475). International Symposium on Environment Software System, James Madison Univ, Harrisonburg, VA, MAY 18–21, 2004.
- Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsas, D., Anthemidis, A., Sofoniou, M., & Kouimtzi, T. (2003). Assessment of the surface water quality in Northern Greece. *Water Research*, 37(17), 4119–4124. doi:10.1016/S0043-1354(03)00398-1.
- Stanimirova, I., Daszykowski, M., & Walczak, B. (2007). Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1), 172–178. doi:10.1016/j.talanta.2006.10.011.
- Vega, M., Pardo, R., Barrado, E., & Deban, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*, 32(12), 3581–3592. doi:10.1016/S0043-1354(98)00138-9.
- Wang, X., Cai, Q., Ye, L., & Qu, X. (2012). Evaluation of spatial and temporal variation in stream water quality by multivariate statistical techniques: a case study of the Xiangxi River basin, Chin. *Quaternary International*, 282, 137–144. doi:10.1016/j.quaint.2012.05.015.
- Weilguni, H., & Humpesch, U. (1999). Long-term trends of physical, chemical and biological variables in the River Danube 1957–1995: a statistical approach, 61(3), 234–259. doi:10.1007/PL00001325.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52. doi:10.1016/0169-7439(87)80084-9.