Analysis of multivariate observations from a monitoring station of a river basin

C. Capilla
Department of Applied Statistics and Operations Research and Quality, Polytechnic University of Valencia, Spain

Abstract

Water quality evaluation provides important information to estimate water system status and to test for compliance with standards. The present work analyzes changes concerning the quality of surface water using a data set from a monitoring station of a Mediterranean river basin. The data set has a dependency structure that renders it multivariate. Principal component analysis is applied to characterize associations present in the multivariate measurements. The principal component scores exhibit temporal correlation. A combined Shewhart-CUSUM control chart is applied to the residuals of the scores time series model to detect changes in the mean level of the data set during the study period. This method detects an outlying observation in the study period due to an extreme value in magnesium concentration. The multivariate assessment of trend is performed using non-parametric tests. The covariance inversion test supported rejection of the hypothesis of no trend in the variables defined with each combination of water quality parameters and month. There is heterogeneity between the trends in the different combinations and an overall trend is not representative. The partial Mann-Kendall test is employed to analyze the trends of each physicochemical variable in the study months. Conductivity trends in two months (May and June) are significant and upward. In the same months calcium trend is also significant but downward. Sodium concentrations exhibit a significant decreasing trend in April. Magnesium levels significantly decrease in March but have an upward trend in June.

Keywords: water quality, multivariate observations, statistical analysis, principal component analysis, control chart, non-parametric trend test.

WIT Transactions on Ecology and The Environment, Vol 175, © 2013 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/ECO130271

1 Introduction

The quality of surface water is a significant concern issue worldwide. Surface waters are degraded by the impact of anthropogenic activities (e.g. agricultural, industrial and urban influences, increasing consumption of water resources), as well as natural processes (e.g. changes in precipitation input, soil erosion). This problem has led administrations to introduce plans and regulations to reduce water pollution. The Council Directive 2008/105/EC of 16 December 2008 established environmental quality standards in the field of water policy and management. This directive has been introduced to protect and achieve sustainable development water resource systems in the European Union countries (Cruz *et al.* [3]). It establishes annual averages and maximum allowable concentrations (environmental quality standards) for priority substances and pollutants. The directive has implied that river basin management plans of most of Europe's river basin were submitted to the European Commission in Brussels in 2010.

The implementation of these plans requires an open and continuous dialogue among policy makers, water managers and scientists. The scientific knowledge of water systems implies observation of their status. Consistent monitoring approaches are needed to indicate actual threats to ecological health. Monitoring provides early warning of water resource system changes which may be due to episodic events, environmental stressors or activities impact. These effects frequently modify the central tendency of observed characteristics by a step change, impulse, monotonic trends or non-stationary trends. The detection and assessment of changes allows determining whether the water resource is improving or deteriorating, and help decision-making to introduce corrective actions in order to improve the system. Data obtained during monitoring are multi-dimensional. Because multiple indicators are required to assess water system quality, the problem is statistically a multivariate problem. The multivariate measurements are not mutually exclusive and are often correlated. Therefore the detection of system changes from these data sets needs the application of multivariate statistical methods. Multivariate techniques, by utilizing the information in the correlation structures of a data set, can provide faster and more efficient anomaly detection for quality assurance and more understandable data analysis for decision-making.

In recent years many studies have applied the multivariate technique principal component analysis (PCA) to interpret water quality parameters. Helena *et al.* [5] investigated the evolution of the groundwater composition between two surveys. They applied PCA to process a set of quantitative analytical data from the alluvial aquifer of the Pisuerga River (Spain). Ouyang [12] employed PCA method to evaluate the effectiveness of the surface water quality monitoring network in a river where the evaluated variables were the monitoring stations. He identified monitoring stations that were important in assessing annual variations of river water quality. Mishra and Tripathi [10] used PCA to analyze a large and complex data set obtained during monitoring of the Ganges River in Varanasi, India. They extracted the parameters that were most important for assessing variation in water quality. Ma *et al.* [9] proposed a model combined PCA and information entropy methods to obtain the weights of water quality indicators, and the proposed model was applied to assess the reused water quality of Jinshui River in Zengzhou City (China).

This paper presents the application of PCA combined with control charts to detect changes in a multivariate data set from a monitoring station in a river basin in Spain. Control charts are on-line schemes widely used to monitor processes. In this study the multivariate data show a strong correlation structure and multivariate methods that consider the variables jointly are required. Univariate control charts can be applied to each individual variable but this method is inefficient and can lead to erroneous conclusions. They may incorrectly identify out-of-control situations which are really in control when analyzed multivariately or may not detect real out-of-control assignable causes (Montgomery [11]). Schall and Chandra [14] developed a method of using PCA to monitor a process which has many characteristics affecting the quality. They applied control charts to PC scores to detect process changes. In this paper the data exhibit autocorrelation and the resulting PC scores are also autocorrelated. Autocorrelation may lead to more frequent false alarms or to slower detection of out-of-control situations and, therefore, to poor performance of control charts (Harris and Ross [4]). When autocorrelation is present one possible adjustment is to model the time dependency with an appropriate ARIMA model (Box *et al.* [2]) and to apply the control chart to the one-step ahead forecast residuals (e.g. Harris and Ross [4]). In this paper the monitoring control scheme is based on the residuals of the PC scores.

Several methods of testing for multivariate trend have been discussed in the water quality literature. In the present article the multivariate assessment of trends is performed using three non-parametric tests. The covariance inversion test is applied to analyze the hypothesis of any monthly trend in any of the observed characteristics (Loftis *et al.* [7]). This work also uses a non-parametric method for testing the homogeneity of trends under dependence of seasons (Smith *et al.* [15]). The partial Mann-Kendall test is also employed to study the trend in each variable per month after removing the effect of the other variables in the same month.

The plan of the remainder of the paper is as follows. Section 2 describes the sampling site, the data used and the statistical methods employed in the analysis. The empirical results and discussion are presented in Section 3. Section 4 contains some concluding remarks.

2 Material and methods

2.1 Sampling site and data set

The study data set was collected from the Ebro River basin in the sampling site of Tortosa (Spain). The Ebro River basin is located in the NE of the Iberian Peninsula, occupying a total surface of 85,362 Km^2. It is the largest hydrographic basin in Spain, accounting for 17.3% of its total surface area. This river discharges in a delta on the Mediterranean Sea at Tortosa, 180 km south of Barcelona. The Ebro is a typical Mediterranean river characterized by seasonal low flows and extreme flush effects, with important agricultural and industrial activity that has caused heavy contamination problems. Water quality concerns in the Ebro basin are salinization processes by natural sources as well as by the pollution produced by irrigation. The biogeochemical characteristics of the river water are highly influenced by anthropogenic activities.The present legislation requires that the water near urban areas be monitored to determine compliance of ecological standards of indicators parameters such as conductance and ions concentrations. The Confederación Hidrográfica del Ebro (CHE) is a government agency established in 1926 to manage the water resources in the Ebro River basin (http://www.chebro.es). CHE has implemented a water quality monitoring network. Monthly data are available on the web page of this agency. In this study the monitoring station located in Tortosa was selected for analysis.

The data set has 4 physicochemical parameters- Conductivity at 20ºC (µS/cm), Sodium, total (mg/L as Na), Magnesium, total (mg/L as Mg) and Calcium, total (mg/L as Ca) - taken on a monthly basis for 14 years (1987-2000). The research focuses on the detection of temporal changes of these water quality indicators. In table 1 the summarized basic statistics of the 4 parameters is presented.

Table 1: Summary basic statistics. Concentration units in mg/L for metal ions. Conductivity in µS/cm.

Variable	Mean	Standard deviation	Minimum	Maximum
Conductivity	946.13	229.95	412	1593
Sodium	82.06	27.04	29.1	139
Magnesium	23.59	6.55	9.7	43.8
Calcium	106.49	16.38	72.1	148.3

2.2 Statistical treatment of the data

In this study the analysis of data was performed using the statistical software R (R Development Core Team [13]). Principal component analysis (PCA) was first applied to analyze the correlation structure of data and to obtain a set of independent parameters for change detection with control charts. PCA is a multivariate statistical technique employed to identify important components that explain most of the variances of a system. For each multivariate measurement (or observation), the PC scores are linear combinations of the standardized p variables. The principal components have two important advantages. The new variables are uncorrelated. A few (sometimes 1 or 2) PCs may capture most of the variability in the data so that we do not have to use all of the p principal components for control. In this work PCA involves the following steps: (1) calculate the correlation matrix of the original data (if the analysis is performed with this matrix the original variables all have equal weights). (2) Find the eigenvalues and the corresponding eigenvectors of the correlation matrix; the eigenvalues of the PCs are a measure of their associated variance, the participation of the original variables in the PCs is given by the components of the eigenvectors (loadings). (3) Discard any components that only account for a small proportion of the variation in the data set. (4) Obtain the score matrix of the remaining PCs. The scores are computed as weighted linear combinations of the standardized original variables where the coefficients are the eigenvector. A detail formalization of PCA method can be found in Jackson [6].

The detection of changes in the mean value of the water quality indicators can be performed by employing control charts to monitor the PC scores (Schall and Chandra [14]). The American Society for Testing and Materials (ASTM) [1] recommends using combined Shewhart-CUSUM charts to detect both sudden large and smaller gradual changes in water quality parameters. Lucas [8] presented a combined Shewhart-CUSUM quality control scheme which includes Shewhart control limits in the CUSUM control procedure. The procedure can signal an out-of-control situation at any stage. The scheme is found to be more sensitive than the CUSUM chart and the Shewhart scheme for small and moderate shifts in the process mean. Application of the Shewhart-CUSUM control chart involves selection of three parameters: h (the decision interval value or value against which the cumulative sum will be compared), k (reference value: a parameter related to the mean change that should be quickly detected), and L (the Shewhart control limit, which is the number of standard deviation units for an immediate release). Previous research (Lucas [8]) showed that k=1, h=5, and L=4.5 were most appropriate for ground water monitoring applications. ASTM [1] suggested the use of h=L=4.5, which results in a single limit with no compromise in leak detection capabilities. With these values the combined control chart quickly detects a mean change of two standard deviations. The early recommended values h=L=4.5 and k=1 are used in this work.

The Shewhart-CUSUM combined method assumes that the observations are independent. In this study the original water quality measurements are autocorrelated. The resulting PC scores also exhibit temporal dependence following an ARIMA seasonal multiplicative model:

$$\phi_p(B)\Phi_P(B) \quad {}^d \quad {}_s{}^D Z_t = \theta_q(B)\Theta_Q(B)a_t \qquad (1)$$

where Z_t are the PC scores in month t, and a_t are the one-step ahead forecast errors or residuals. A detailed description of the ARIMA modelling approach can be found in Box *et al.* [2]. The assumption of uncorrelated data is critical to the performance of control charts, which are very inefficient when monitoring characteristics are dependent (Harris and Ross [4]; Montgomery [11]). Autocorrelation may lead to more frequent false alarms or to slower detection of out of control situations. In this work the control charts are used to monitor the residuals a_t of the PC scores. The residual-based control charts have been proposed in the literature (Harris and Ross [4]) as an appropriate method to deal with autocorrelation in monitoring measurements. The residuals a_t are due to random perturbations in the process and under control are independently and normally distributed with zero mean and variance σ_a^2.

The Shewhart-CUSUM control chart is an acceptable method for detecting step and impulse mean-level changes of water quality parameters. Other possible out of control situations are the presence of increasing concentrations with time or seasonal trends in data. The detection of these changes can be performed by using methods for the assessment of trends. Water quality data usually are multivariate. They are often correlated in time and/or highly skewed with extreme values. Some of the values may be missing. Some of the above-mentioned problems are present in this study set. Because of these limitations, the use of the residual-based Shewhart-CUSUM control chart as the sole statistical method for trend detection is inappropriate.

Appropriate techniques have been developed in the applied literature that is based on non-parametric statistics to test multivariate trends (Smith *et al.* [15]). These methods are resistant to the problems mentioned above. In this study, trend analysis is performed using three non-parametric multivariate trend tests. The first method is the covariance inversion test (Loftis *et al.* [7]). This technique is implemented with a quadratic form as test statistic using the Kendall's rank correlation between time and variable. The 4 water quality parameters and the 12 months are viewed as defining 48 variables and 48 trend statistics, to test the null hypothesis of no trend. The test statistics follows a chi-square (χ^2) distribution with 48 degrees of freedom under the null hypothesis. The second test is an extension of the covariance inversion test proposed by Smith *et al.* [15] and its objective is to assess homogeneities in trends. The method decomposes the covariance inversion test statistics into two terms. The first term asymptotically displays a χ^2 distribution with one degree of freedom under the null hypothesis of no overall trend. The second term (remainder) is considered to be that which persists in all 48 trends after removing an average trend. Under the null hypothesis of homogeneous trend, it asymptotically follows a χ^2 distribution with 47 degrees of freedom distribution. The insignificance of the remainder validates the use of the first term as a test statistic for analysing an overall trend. The third non-parametric method is the partial Man-Kendall test, which is a conditional test for trend in each characteristic after removing the effect of the other characteristics. We employ this method for each combination of variable and month to test the presence of significant trends.

3 Results and discussion

In PCA the first step is to diagonalize the data correlation matrix. Table 2 displays this matrix for the 4 water quality parameters. The pairwise Pearson's correlation coefficients show that there is a strong correlation structure in data. All the correlations are significantly different from zero (alpha=0.01). The

Table 2: Correlation matrix.

Variable	Conductivity	Sodium	Magnesium	Calcium
Conductivity	1	0.901	0.760	0.771
Sodium	0.901	1	0.809	0.834
Magnesium	0.760	0.809	1	0.730
Calcium	0.771	0.834	0.730	1

number of PCs which can be computed is equal to the number of variables. The first PC accounts for about 84.5% of the total variance of the data. The first two PCs sum almost 91.9% of the total variance in the water quality observations. Table 3 gives the eigenvectors (loadings) for the first and the second components.

Table 3: Loadings of the first and second PC.

Variable	Principal component 1	Principal component 2
Conductivity	0.503	0.173
Sodium	0.523	0.126
Magnesium	0.483	-0.833
Calcium	0.490	0.509

The absolute value of the loadings is an indicator of the participation of the physicochemical variables in the PCs. All variables are well represented by the first PC. This component reflects the positive correlation that exists among them. The variables that primarily contribute to the second component are magnesium and calcium. Their two loadings have different sign revealing than in the sample there are months with high concentrations of calcium and small concentrations of magnesium, and vice versa. The most significant parameter in this second component is magnesium. Its scores can be useful to detect mean-level changes in magnesium and calcium.

The PC scores are computed as weighted lineal combinations of the original water quality observations. The coefficients are the loadings given in table 3. The analysis of the scores of the chosen PCs shows that there is autocorrelation in PC1 and PC 2 scores (fig. 1). The autocorrelation and partial autocorrelation functions in fig. 1 a) and b), suggest that the first PC scores follow an ARIMA seasonal multiplicative model of order $(1,0,0)x(1,0,0)_{12}$:

$$Z_t = \phi Z_{t-1} + \Phi Z_{t-12} + a_t \tag{2}$$

The second PC scores follow an ARIMA model of order (0, 0, 1):

$$Z_t = \theta a_{t-1} + a_t \tag{3}$$

In eqns (2) and (3) Z_t represent PC score at month t and a_t are the one-step ahead forecast errors or residuals, which are independent. The detection of mean-level changes can be performed by simultaneously monitoring the residuals a_t of PC 1 and PC2 time series model. Fig. 2 represents the residual-based Shewhart-CUSUM control charts.

The combined chart for the residuals of PC 1 (Fig.4a)) shows no out-of-control signal. In fig. 2b) the lower cumulative sum and the absolute value of the residuals exceed the control limit (blue line, h=L=4.5) on observation number 152. The association of PC 2 with magnesium and calcium concentrations indicates that there are extreme values in at least one of these two physicochemical parameters. There is an extreme value of magnesium (43.8 mg/L) at time point 152 (August 1999) which implies an outlying multivariate observation.

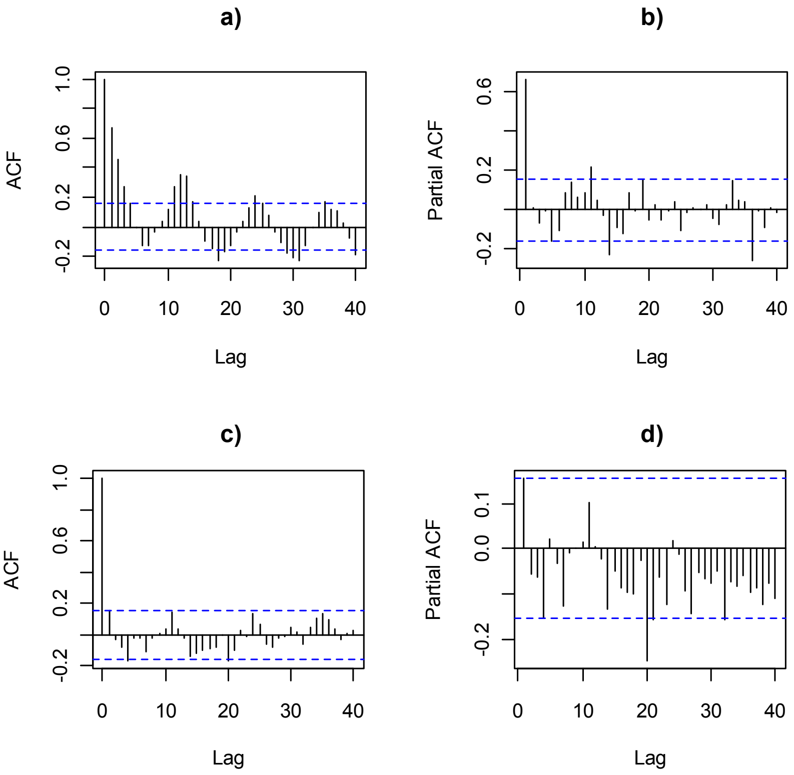

Figure 1: a) Autocorrelation function of the first PC scores. b) Partial autocorrelation function of the first PC scores. c) Autocorrelation function of the second PC scores. d) Partial autocorrelation function of the second PC scores.

Tables 4 and 5 display the results of the nonparametric tests as applied to the water quality data set. First, view the four constituents and twelve months as defining 48 variables and 48 trend statistics. The associated covariance matrix is of dimension 48 x 48. The covariance inversion test (table 4) rejects the null hypothesis of no trend at the 5% level (p-value<0.05). The test statistic is decomposed into two terms (see the third and four rows of table 4).

The remainder is what persists in all 48 trends after the removal of an average trend. This is significant (χ^2=71.6836 with 47 degrees of freedom, P-value <0.05). This indicates that there is heterogeneity between the trends in the different combinations of variables and months, and that an overall trend is not representative.

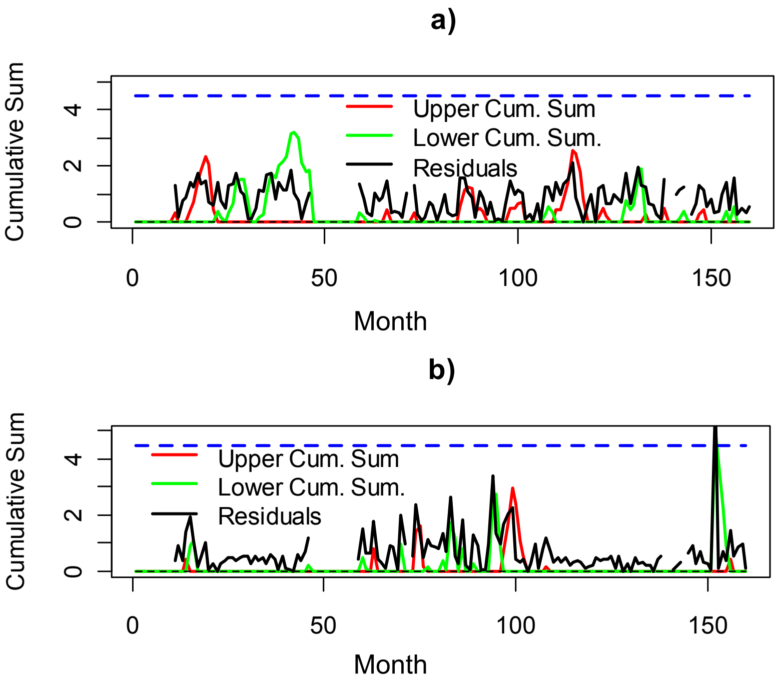

Figure 2: a) Control chart of the residuals of the first PC ARIMA model. b) Control chart of the residuals of the second PC ARIMA model.

Table 4: Results of the multivariate trend tests based on Kendall's rank correlation between time and variable.

View	Test method	Chi-Square	Degrees of freedom	p-value
48 variables (4 water quality parameters x 12 months)	Covariance inversion test	71.7263	48	0.0148
	Decomposition:			
	Overall trend	0.0427	1	0.8362
	Remainder	71.6836	47	0.0117

A detailed analysis of the trend for each variable and for each month is given in Table 5. The method is the partial Mann-Kendall trend test. The values for each month in Table 5 are the conditional trend estimations after removing the effect of the other variables in the same month. The conditional test shows that Conductivity and Calcium have significant trends in late spring and early summer. The trend of Conductivity is upward and of Calcium downward. Sodium and Magnesium have decreasing trends in spring, in April and March respectively. Magnesium also has a significant increasing trend in June. Calcium presents a significant decreasing trend in September.

Table 5: Results of the application of the Partial Mann-Kendall test.

Month	Conductivity	Sodium	Magnesium	Calcium
January	0.009	-0.132	0.229	-1.181
February	-0.269	-0.845	0.454	-0.121
March	-0.688	1.257	-2.149*	0.751
April	1.477	-1.799*	0.722	-0.242
May	2.322*	-0.346	-0.225	-2.304*
June	6.024**	-1.058	1.972*	-2.153*
July	1.129	-0.703	1.272	-0.769
August	-0.035	-0.405	1.606	0.119
September	0.935	1.253	0.145	-1.687*
October	0.822	-0.076	-0.875	-0.144
November	0.510	1.112	-0.589	-0.347
December	1.112	1.203	-0.327	-0.538

*Significant at the 5% level.
**Significant at the 1% level.

Multiple factors could have been affecting the river water quality trends. An attempt should be made to correlate the observed trends with activity logs for this river basin. The overexploitation and irrigation return flow has resulted in the significant increase of magnesium in June. Conductivity upward trend is significant in May-June because in this period the body of water has more dissolved materials such as nutrients and salts, which indicates poor water quality. Sources of total dissolved nutrients can include agricultural runoff, domestic runoff or discharges of wastewater treatment plants. The significant decrease in April accounts for a decrease in salinity in this month.

4 Conclusions

In this study changes in surface water quality in a monitoring site of Ebro River basin (Spain) have been assessed. For each sample, four physicochemical parameters have been considered: Conductivity, Sodium, Magnesium and Calcium. The data have been taken on a monthly basis for 14 years (1987–2000). The observations have been processed by multivariate statistical techniques in order to investigate the presence of changes which may be due to episodic events, environmental stressors or anthropogenic activities impact. The inspection of the correlation matrix of the four analyzed variables showed the existence of strong correlations between them. The application of principal component analysis offered a better understanding of water quality status of the studied system. This method revealed some specific features of the data structure. Two principal components were obtained summing almost 92% of the total variance (i.e. information) in the water data set. The first component had similar loadings for all the water quality parameters and represents the strong positive correlation among them. The second component was linked with magnesium and to a lesser extent with calcium, and measured the difference among these two characteristics.

Autocorrelation and partial autocorrelation plots of the two first components showed that principal component scores were not independent when viewed over time. The scores were treated using standard time series modeling approach. An ARIMA model was estimated and allowed to obtain the one-step ahead forecast errors or residuals. The residuals were independent. The application of the Shewhart-CUSUM control chart to the residuals detected an outlying observation in the study period due to an extreme value in magnesium concentration. Ignoring the serial correlation when it exists can severely impair the interpretation of control charts.

Multivariate non-parametric trend tests were employed to detect trends. The covariance inversion test supported rejection of the hypothesis of no trend in the 48 variables defined with each combination of variable and month. A decomposition of this test statistic showed that the trends are significantly different across variables and seasons. Trends for each combination of variable and month were analyzed using the partial Mann-Kendall test. Conductivity trends in two (May and June) of twelve months were significant and upward. In the same seasons calcium trend was also significant but downward. Sodium concentrations exhibited a significant decreasing trend in April. Magnesium levels significantly decreased in March but had an upward trend in June. These results could be useful to the water quality managers and policy –makers to select optimal pollutant reduction strategies, to avoid exceeding environmental quality standards, and to introduce corrective actions in order to improve the system.

References

[1] ASTM, *Standard Guide for Developing Appropriate Statistical Approaches for Ground Water Detection Monitoring Programs,* ASTM D6312, West Coshohocken, Pennsylvania, American Society of Testing and Materials, pp. 34-50, 2005.

[2] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C., *Times Series Analysis Forecasting and Control*, 3rd edition, Prentice Hall: Englewood Cliffs, New Jersey, pp. 327-370,1994.

[3] Cruz, J.V., Pacheco, D., Coutinho, R., Cymbron, R., Mendes, S., Antunes, P., Fonteila, J. and Freire, P., Chemical monitoring of river water bodies in an EU outmost region: examples from the Azores archipelago (Portugal). *Journal of Environmental Monitoring*, **12(12)**, pp. 2216-2225, 2010.

[4] Harris, T.J. and Ross, W.H., Statistical process control procedures for correlated observations, *The Canadian Journal of Chemical Engineering*, **69(1)**, pp. 48-57, 1991.

[5] Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M. and Fernandez, L., Temporal evolution of groundwater composition in an

alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research*, **34(3)**, pp. 807-816, 2000.

[6] Jackson, J.E., *A User's Guide to Principal Component Analysis*, Wiley Interscience: Hoboken, New Jersey, pp. 569, 2003.

[7] Loftis, J.C., Taylor, C.H. and Chapman, P.L., Multivariate test for trend in water quality. *Water Resources Research*, **27(7)**, pp. 1421-1429, 1991.

[8] Lucas, J.M., Combined Shewhart-CUSUM quality control schemes. *Journal of Quality Technology*, **14(2)**, pp. 51-59, 1982.

[9] Ma. J., Guo, J. and Liu, X., Water quality evaluation model based on principal component analysis and information entropy: application in Jinshui river, *Journal of Resources and Ecology*, **1(3)**, pp. 249-252, 2010.

[10] Misha, A. and Tripathi, B.D., Ecological investigation of the Ganges River using principal component analysis. *Environmental Quality Management*, **19(1)**, pp. 61-70, 2009.

[11] Montgomery, D.C., *Introduction to Statistical Quality Control*, 5th edition, John Wiley & Sons: New York, 2005.

[12] Ouyang. Y., Evaluation of river quality monitoring stations using principal component analysis, Water Research, 39(12), pp. 2621-2635, 2005.

[13] R Development Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-90051-07-0, http;//CRAN.R-project.org, 2012.

[14] Schall, S. and Chandra. J., Multivariate quality control using principal components, *International Journal of Production Research*, **25(4)**, pp. 571-588, 1987.

[15] Smith, E.P., Rheem, S. and Holtzman, G.I., Multivariate assessment of trends in environmental variables (Chapter 23). *Multivariate Environmental Statistics*, ed. G.P. Patil and C.R. Rao, Elsevier Science Publishers: The Netherlands, pp. 489-507, 1993.