INTERNATIONAL JOURNAL OF SCIENCE AND NATURE

© 2004 - 2017 Society For Science and Nature(SFSN). All Rights Reserved

www.scienceandnature.org

PRINCIPAL COMPONENT ANALYSIS (PCA) FOR THE ASSESSMENT OF HYDROLOGICAL AND PHYSICOCHEMICAL PARAMETERS OF RIVER WATER

MHJP Gunarathna[1], AMKR Bandara[2], MKN Kumari[1] and GY Jayasinghe[3]

[1]Department of Agricultural Engineering & Soil Science, Faculty of Agriculture, Rajarata University of Sri Lanka,
Puliyankulama, Anuradhapura, Sri Lanka.
[2]Department of Agricultural Systems, Faculty of Agriculture, Rajarata University of Sri Lanka, Puliyankulama,
Anuradhapura, Sri Lanka.
[3]Department of Agricultural Engineering, Faculty of Agriculture, University of Ruhuna, Mapalana,
Kamburupitiya, Sri Lanka.
*Corresponding authors email: janaka78@gmail.com

ABSTRACT

Physicochemical and hydrological parameters of *Yan* River and land use pattern of *Yan* River basin were assessed by principal component analysis (PCA). Water samples were collected from seven locations along the *Yan* River once a month during 2014/15, representing pre monsoonal, monsoonal and post monsoonal periods to estimate thirteen physicochemical parameters. Seven parameters related to land use and five hydrological parameters representing rainfall and discharge were also used for this study. The PCA produced four main components (*viz*. mineral dissolution, miscellaneous factors, land use related factors and factors related to nitrogen concentration) and explained more than 72.37% of the variance in water quality. The first two verymax factors (VFs) explained 43.46% of the total variance. Minerals showed negative loading, in contrast to positive loading of total discharge and monthly rainfall of the upper catchment area. The monthly scores suggested separating the months into two clusters, as monsoonal and post monsoonal periods. Third and fourth VFs explained 28.91% of the total variance. Agricultural lands and homesteads showed positive loading, in contrast to the water bodies in sub-watershed land use. Percentage of agricultural lands and homesteads in sub-watershed shows positive loading on to VF4 with NH_4-N and NO_3-N, alarming that substantial amounts of nitrogen fertilizer applied to agricultural lands tend to wash off and cause the water pollution. Percentage of marshy lands negatively loaded on to VF4, indicating that high uptake and assimilation of dissolved nitrogen in water by the vegetation in marshy lands.

KEYWORDS: hydrological, land use, physicochemical, principal component analysis, variance.

INTRODUCTION

Water quality of rivers are mainly affected by natural processes including dissolution of minerals from overlying rocks, anthropogenic activities such as agricultural, urban and industrial and enhanced natural processes due to human behaviors such as erosion and climate change. Since water quality has a great influence on aquatic lives and water availability for human and other animals, it is imperative to monitor the water resources for their sustainable use (Yang *et al*., 2009; Oketola *et al*., 2013). Further, the reliability of assessments should be high since rivers are playing a significant role in water availability. When monitoring of river water quality, it is essential to assess the physicochemical parameters of water on both temporal and spatial basis (Iqbal *et al*., 2004). Consideration of hydrological parameters and land use patterns of sub watersheds are also vital in assessing water quality (Wang *et al*., 2014). With a higher number of physicochemical parameters and sampling points along the river, it will increase the size of data matrix which will lead to complex situations in the data analysis and interpretation of results (Chapman, 1992). Many authors have used different techniques to analyze data, including data reduction techniques to simplify the interpretation process of large data sets (Yang *et al*., 2009; Fan *et al*., 2010; Thareja and Trivedi, 2010; Oketola *et al*., 2013). These data reduction techniques have merits and demerits, which are unique to each method. Many authors paid attention on water quality indices on decision making using large data sets. It has demerits, unique to pollution type and geographical areas, therefore universal application is limited. Univariate is also commonly used in water quality analysis; however it is not capable enough to characterize similarities and differences of variables. Multivariate analysis techniques show wide applications in data reduction of water quality studies with the merit of characterization of similarities and differences, which gives good clues on source of pollution, *etc*. Principal component analysis is a multivariate analysis method, which shows relatively superior results in data reduction, while showing association of inter related variables, which leads to draw meaningful conclusions using large data matrices (Yang *et al*., 2009; Fan *et al*., 2010; Thareja *et al*., 2011). *Yan* River originates from the hilly areas of *Dambulla* and *Sigiriya* and flows towards the North-Eastern region of Sri Lanka. It has a catchment area of 1538 km^2 which is located entirely in the dry zone of Sri Lanka. *Yan* River fulfills the water requirement of nearby residents including their irrigation requirements. Further, it discharges about 482 MCM to sea annually, mostly contributed by the second inter monsoonal and north east monsoonal rains (Manchanayake and Madduma Bandara, 1999). Small streams and tributaries that pass by agricultural lands (mainly paddy fields) and homesteads enter the *Yan* River in several places and this may affect the quality of water, further. This study was aimed to assess physicochemical variation of *Yan* River with respect to the hydrological conditions and land use pattern in *Yan* River basin by using principal component analysis (PCA).

MATERIALS & METHODS

Sampling points

Water samples from seven locations along the middle of *Yan Oya* were collected (Figure 01). Characteristics of land use, major water inputs and special features of sampling points are listed in the Table 1.

TABLE 1: Sampling locations and characteristics of respective sub watersheds

Sampling point number	Remarks
01	Located 3 km away from the *Huruluwewa* reservoir Major water input is drainage water of paddy fields of the commanding area of *Huruluwewa* reservoir It also received spilled water during the spilling period of the reservoir
02	Located 3.5 km north from the sampling point 01 Major water input is drainage water of paddy fields of a tank cascade system and drainage water of homestead areas of *Yakalla* town
03	Located 4.5 km north from the sampling point 02 Major water input is drainage water of paddy fields and drainage water of homestead areas of *Galenbindunuwewa* town
04	Located 6 km north east from the sampling point 03 Major water input is drainage water of paddy fields and shrub lands
05	Located 4.5 km north east from the sampling point 04 Major water input is drainage water of paddy fields of a tank cascade system and drainage water of homestead areas of small villages
06	Located 11.5 km north east from the sampling point 05 Major water input is drainage water of paddy fields and drainage water of homestead areas of small villages
07	Located 14 km away north east the sampling point 06 Major water input is drainage water of paddy fields and drainage water of homestead areas of small villages

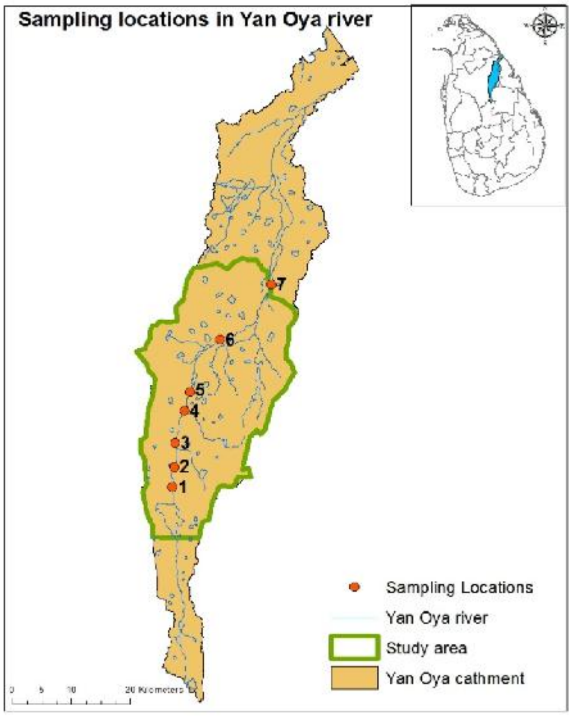

FIGURE 1: Study area and sampling locations

Hydrological data

Rainfall data of *Huruluwewa* gauging station (which represent the upper catchment) and *Horowpothana* gauging station (which represents the lower part of the catchment) were gathered from the Meteorological Department of Sri Lanka. Based on the available data, average rainfall over the catchment was estimated by using arithmetic mean method.

Stream flows of seven sampling locations were measured using velocity area method. Flow velocities were measured by using FLOWATCH JDC 94956 flow velocity meters at the depths of 0.2 and 0.8 water heights. Cross section of the river in each sampling points were drawn and the areas, respective to the water heights were calculated. The discharge of each sampling locations was calculated using average velocity measured and area estimated. Four hydrological variables such as monthly rainfall (mm) of *Huruluwewa* (MRF_Hu), and *Horowpothana* (MRF_Ho), average monthly rainfall (mm) of the catchment (Av_MRF) and the discharge (m^3/s) at respective sampling points (Q) during the water sampling were used for the analysis.

Land use data

Land use data of year 2014 were obtained from the Department of Land Use and Planning of Sri Lanka and land use at sub watershed level were extracted. Sub watershed areas of respective sampling locations were delineated using hydro tool of ArcMap 10.2.1 software of ESRI. Elevation data derived from Google earth pro software was used to produce contours at 5 m interval, which is required in the delineation process. Land uses were categorized into four major groups as agricultural and homestead (paddy, chena, upland perennials and plantations), forests and scrubs (forests, forest plantations, scrubs and marshy lands), water bodies (tanks, ponds, streams, *etc.*) and others (gravel pits, roads, rocks, *etc.*) Areas of each land use group in sub watersheds respective to the sampling points were estimated and percentage of existing agricultural and homestead, forests and scrubs and water bodies were calculated. Further, percentages of existing aforesaid land uses over the upper watershed areas of respective sampling points were also estimated. Percentage of six land use pattern variables as percentage of agricultural and homesteads in sub watershed (SWLU_AH%), percentage of forests and scrubs in sub watershed (SWLU_F%), percentage of water bodies in sub watershed (SWLU_W%), percentage of agricultural and homesteads in the watershed above the respective sampling location (WLU_AH%), percentage of forests and scrubs in the watershed above the respective sampling location (WLU_F%), percentage of water bodies in the watershed above the respective sampling location (WLU_W%), were used for the analysis. Further, the percentages of marshy lands above the sampling point within a half circle of 500 m radius (ML) were also used as the seventh land use parameter for the analysis.

Physicochemical analysis of water

Water sampling was done in 2014/15 at monthly intervals to estimate 13 physicochemical parameters of water. At the time of sampling, total dissolved solids (TDS), electrical conductivity (EC), dissolved oxygen (DO), pH and temperature were measured by using instruments mentioned in Table 2. The water samples were collected, transported and stored for further analysis in soil and water laboratory of the Department of Agricultural Engineering and Soil Science, Faculty of Agriculture, Rajarata University of Sri Lanka. Procedures explained in APHA guidelines were followed for the examination of water and wastewater (APHA, 1992). Available nitrate nitrogen, ammoniacal nitrogen, calcium, magnesium, sodium, potassium concentrations and total suspended solids (TSS) were measured using methods listed in the Table 2 and the procedures explained in APHA guidelines. Sodium absorption ratio (SAR) was calculated by using measured data.

TABLE 2: Instruments and methods used for water quality analysis

Water quality parameter	Instrument / Method
Temperature (T)	DO meter (EUTECH, CyberScan DO 300)
Dissolved Oxygen (DO)	DO meter (EUTECH, CyberScan DO 300)
Turbidity (TBD)	Turbidity meter (EUTECH, TN 100)
pH	Multi parameter analyzer (HATCH, Sension 156)
Electrical Conductivity (EC)	Multi parameter analyzer (HATCH, Sension 156)
Total dissolved solids (TDS)	Multi parameter analyzer (HATCH, Sension 156)
Ammoniacal nitrogen (NH_4–N)	4500 NH_3 F Phenate method
Nitrate nitrogen (NO_3 –N)	Salicylic Acid method
Ca^{2+} concentration (Ca)	Atomic absorption spectrophotometer
Mg^{2+} concentration (Mg)	Atomic absorption spectrophotometer
Na^+ concentration (Na)	Flame photometer
K^+ concentration (K)	Flame photometer
Total suspended solids (TSS)	Oven drying method

Principal component analysis (PCA)

Initially PCA was carried out using all 24 variables, consisting of thirteen physicochemical, four hydrological and seven land use related, as input variables. Since these variables were measured on widely different scales, PCA was done based on correlation matrix. The suitability of dataset for PCA was assessed by calculating the correlation coefficients between variables, determinant of correlation matrix, Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (Kaiser,1974) and Bartlett's test of sphericity (Bartlett, 1954). The multi-coliniarity problem was solved by removing highly correlated variables (r > 0.8) one by one, until the determinant value of the correlation matrix reached up to the critical value 0.00001 (Field, 2005). The variables with small values on the diagonals of the anti-image correlation matrix were removed until overall KMO statistics reached the acceptable level, 0.5 (Hutcheson and Sofroniou, 1999). After eliminating the problematic variables, an unrotated PCA was performed to identify the significant principal components (PCs). The ideal number of principal components was determined using the Kaiser criterion (eigenvalue > 1 rule) (Kaiser, 1960) and Scree test (Cattell, 1964). A new PCA was performed with varimax rotation and the selected number of components in order to facilitate the interpretation (Abdi, 2003). Varimax factor (VF) coefficient having a correlation >0.75 were regarded as strong significant factor loading (Liu *et al*., 2003). Meanwhile VF in the range of 0.75-0.50 and 0.50-0.30 were considered as moderate and weak factor loading, respectively. Problematic variables were identified by performing PCA with SPSS software (version, 22.0) and final PCA with varimax rotation was done using XLSTAT 2014.

RESULTS & DISCUSSION

Determinant value of correlation matrix of all the variables under study was found 8.118 x 10^{-11} indicating high multi-collinearity that exists between the variables. For the purpose of improving the determinant value, seven variables with correlation coefficients (r) > 0.8 were removed. After removing them, the determinant value was worked out to be 0.0000109, which is greater than the threshold value. By this approach, the problem of high multi-collinearity was removed and the rest of the variables were moderately correlated with each other. The value of Kaiser-Meyer-Olkin (KMO) statistics was 0.457, which is less than the bench mark (0.5) and it indicates that the sample is not adequate to perform PCA. For improving the KMO statistics, the variables with the lowest values on diagonals of the anti-image correlation matrix was removed to achieve a KMO statistics of 0.5724. It showed that the sample size was adequate for PCA. On the other hand, the Bartlett's test of sphericity was statistically significant (p < 0.0001) and it indicated the factorability of the correlation matrix. This implied the suitability of applying the PCA for finally selected set of variables.

Table 03, represents the determined initial PCs, their eigen values and percent of variance contributed by each PC. Figure 02, shows the scree plot of the eigen values and the percentage of cumulative variability explained by each component. The eigen value greater than one was taken as criterion for the extraction of principal components required to explain the sources of variances in the data (Iscen *et al*., 2008).

TABLE 3: Initial components

Component	Eigen value	Variability (%)	Cumulative %
1	4.5554	28.4712	28.4712
2	2.6512	16.5701	45.0414
3	2.2642	14.1510	59.1924
4	2.1087	13.1796	72.3720
5	0.9897	6.1859	78.5579
6	0.8772	5.4826	84.0405

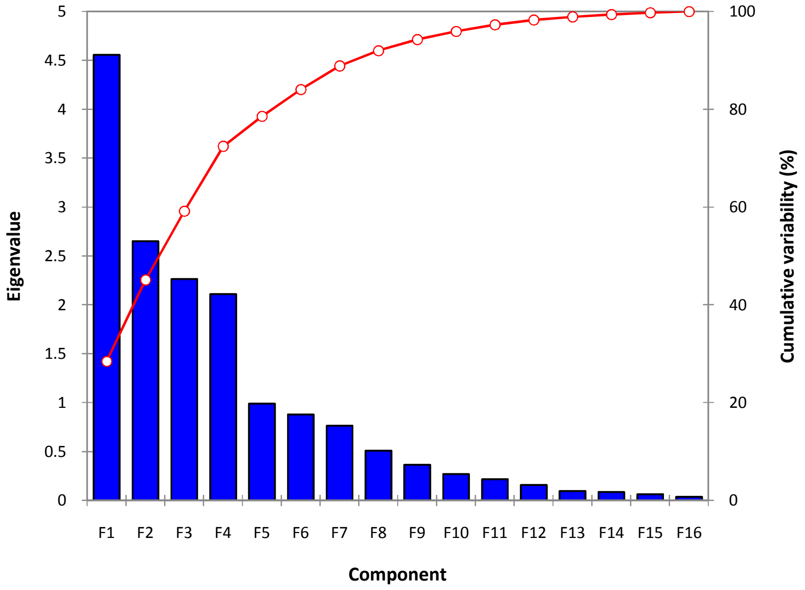

FIGURE 2: The scree plot

As shown in the Figure 02, the inflection point occurred at the component sequence number of four after which the eigen values were less than one. Thus, four components were extracted and Table 03 shows all four components, which explains 72.37% of the total variability. Before rotation, first component accounted for 28.47% of total variance than the remaining three components (16.57%, 14.15% and 13.17%, respectively).

Table 4 shows the orthogonal rotated component matrix of variables, which reflects the correlation coefficients between the variables under study and four components. The first component has higher loading of the six variables related to mineral dissolution. This includes Ca, Mg and Na concentration, TDS, monthly rainfall at *Huruluwewa* (MRF_Hu) and total discharge (Q).

TABLE 4: Rotated Component Matrix of Variables (Factor loadings after Varimax rotation)

Variable	VF1	VF2	VF3	VF4
Ca	-0.6753	0.5849	0.1565	-0.1319
Mg	-0.8646	-0.1163	-0.0560	-0.0863
Na	0.5245	0.2018	-0.0987	0.1729
K	0.0093	-0.7871	-0.1335	0.0630
NH4-N	0.0945	-0.0731	0.0589	0.8584
NO3-N	0.3055	0.1189	0.0703	0.6470
DO	0.4244	0.8513	-0.0734	0.1005
TDS	-0.8031	-0.2538	-0.0122	-0.0995
pH	-0.4330	0.5694	-0.0443	-0.3807
T	-0.2944	-0.6377	0.1250	-0.2012
MRF_Hu	0.9262	-0.1799	-0.0006	0.1244
Q	0.8832	0.0996	0.0738	-0.2383
SWLU_AH%	-0.1014	-0.0285	0.7228	0.6088
SWLU_W%	-0.0023	-0.0571	-0.7521	0.1342
WLU_AH%	0.0182	-0.0363	0.9272	-0.0035
ML	0.1483	-0.0186	0.3963	-0.7402

The second component has higher loading of K concentration, DO, pH and temperature. The third component has higher loading of the three variables related to land use. This includes SWLU_AH%, SWLU_W% and WLU_AH%. Finally, the fourth component has higher loading of the four variables related to nitrogen content. This includes NH4-N, NO3-N, SWLU_AH% and ML.

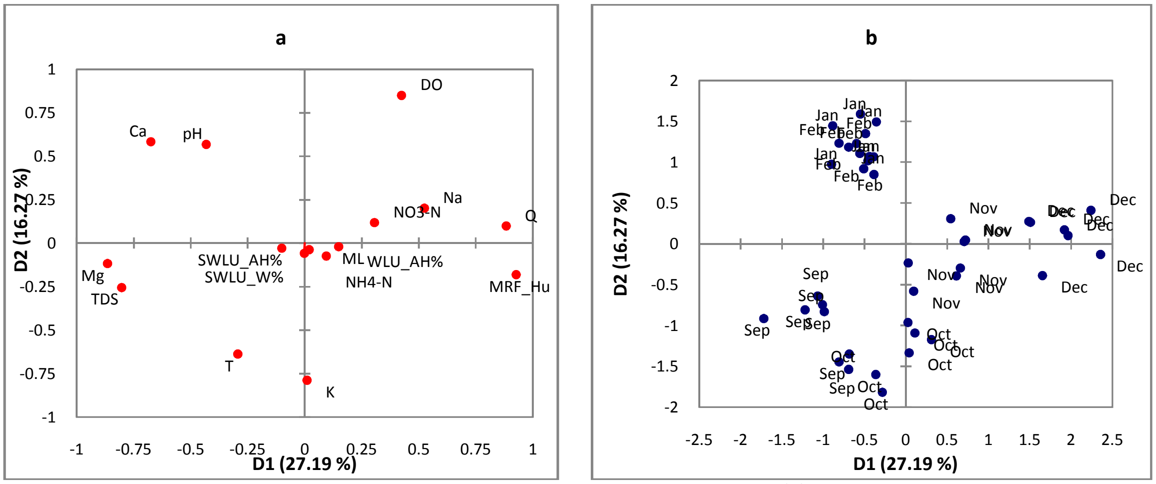

FIGURE 3: PCA loading (a) and Monthly scores (b) of first two VFs

The loadings and the monthly scores of the first two VFs (VF1 vs. VF2) are presented in Figure 03. The first two VFs explained 43.46% of the total variance. Mg and TDS showed higher negative loadings and Ca showed moderate negative loading, which is in contrast to higher positive loading of monthly rainfall of *Huruluwewa* (MRF_HU) and total discharge (Q) on VF1. This contradiction may occur due to reduction of dissolved solid concentrations as the volume of water increases. The monthly scores plot (Fig. 3b) suggests the separation of months into two clusters; first cluster consists of January and February and the second cluster consists of months from September to December. These two clusters can be interpreted as post monsoon period and monsoonal rainy period, respectively. The interesting feature in this plot is that the ordering of months in cluster two alone the VF1 with the lowest score for September and the highest score for December. This pattern is same as the ascending order of the monthly rainfall received in the Dry zone of Sri Lanka. Figure 4 shows the PCA loadings and the location score of third and fourth VFs (VF3 vs. VF4). These two VFs explained 28.91% of the total variance. Agricultural lands and homesteads above the sampling point (WLU_AH%) and sub-watershed (SWLU_AH%) showed a higher positive loading on VF3, whereas extents of water bodies in sub-watershed land-use (SWLU_W%) showed a higher negative loading onto VF3.

This may be due to less availability of land for agriculture when there is a higher proportion of a water body in a sub-watershed. Further, the percentage of agricultural lands in sub-watershed (SWLU_AH%) showed positive loading on to VF4 with NH4-N and NO3-N. This indicates that substantial amount of nitrogen fertilizer applied to agricultural lands tend to wash off and leading to water pollution. On the other hand a percentage of marshy land (ML) negatively loaded on to VF4 indicated that there is a negative correlation between the amount of nitrogen and the percentage of marshy lands, as decreasing of nitrogen content in water with increasing of marshy lands. This may be due to the uptake and assimilation of dissolved nitrogen in the water by vegetation in the marshy land. This conclusion is also supported by not loading the percentage of agricultural land in the whole watershed (WLU_AH%) on to VF4, while only a percentage of agricultural land in the sub watershed was positively loaded, indicating that a percentage of agricultural lands in the whole watershed (WLU_AH%) is not affected due to the amount of dissolved nitrogen.

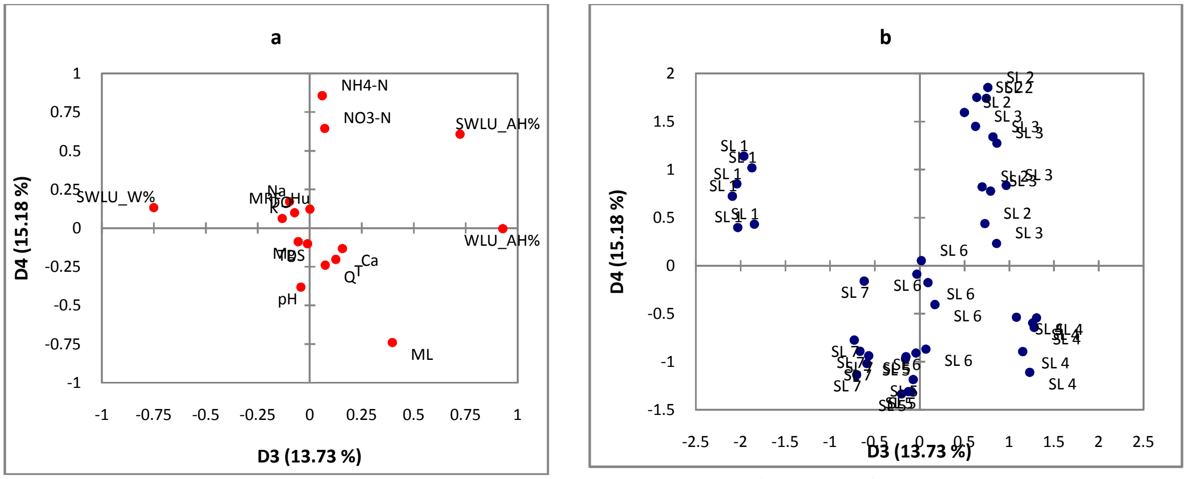

FIGURE 4: PCA loading (a) and location scores (b) of third and fourth VFs

CONCLUSION

The PCA produced four main components (*viz.* mineral dissolution, miscellaneous factors, land use related factors and factors related to the nitrogen concentration) and explained more than 72.37% of the variance of water quality.

The first two VFs explained 43.46% of the total variance. Minerals showed a negative loading in contrast to positive loading of total discharge and monthly rainfall of upper catchment area. The monthly scores suggested separating the months into two clusters, as monsoonal and posting monsoonal period.

Third and fourth VFs explained 28.91% of the total variance. Agricultural lands and homesteads showed positive loading in contrast to the water bodies in the sub-watershed land use. Percentage of agricultural lands and homesteads in sub-watershed showed a positive loading on to VF4 with NH_4-N and NO_3-N, alarming that substantial amount of nitrogen fertilizer applied to agricultural lands tends to wash off causing water pollution. A percentage of marshy lands negatively loaded on to VF4 and it indicated that the dissolved nitrogen in water could be up taken and assimilated in high amounts by their vegetation. PCA can be successfully used to understand the variations of physicochemical parameters of water in contrast to the hydrological parameters and land uses.

ACKNOWLEDGMENTS

Authors wish to acknowledge the financial assistance given by Research, Publication and Higher Degrees Committee of Rajarata University of Sri Lanka for the success of this research work. Further, authors wish to extend sincere gratitude to the Department of Irrigation, Department of Meteorology and Department of Land Use and Planning for providing necessary data for the success of this Study.

REFERENCES

Abdi, H. (2003) Multivariate analysis. In M. Lewis-Beck, Bryman A. and Futing, T. (Eds), *Encyclopedia for research methods for the social sciences*. Thousand Oaks, CA: Sage.

American Public Health Association (1992) *Standard methods for the examination of water and wastewater*, 18th edition. America Public Health Association, Washington.

Bartlett, M.S. (1954) A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society* 16: 296–298.

Cattell, R.B. (1966) The scree test for the number of factors. *Multivariate Behavioral Research* 1(2): 245-76.

Chapman, D. (1992) Water Quality Assessment. 2nd Edition. World Health Organization, New York, USA.

Fan, X., Cui, B., Zhao, H., Zhang, Z. and Zhang, H. (2010) Assessment of river water quality in Pearl River Delta using multivariate statistical techniques. *Procedia Environmental Sciences* 2: 1220–1234.

Field, A. (2005) *Discovering Statistics Using SPSS.* 2nd edition, SAGE, London.

Hutcheson, G., and Sofroniou, N. (1999) *The multivariate social scientist: introductory statistics using generalized linear models.* London: Sage Publication.

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. *Environmental Monitoring and Assessment* 144(1-3): 269-276. http://dx.doi.org/10.1007/s10661-007-9989-3

Iqbal, F., Ali, M., Salam, A., Khan, B. A., Ahmad, S., Qamar, M. and Umer, K. (2004) Seasonal Variations of Physico-Chemical Characteristics of River Soan Water at Dhoak Pathan Bridge (Chakwal), Pakistan. *International Journal of Agriculture & Biology* 6(1): 89 – 92.

Kaiser, H. F. (1960) The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20: 141-51.

Kaiser, H. (1974) An index of factorial simplicity. *Psychometrika* 39: 31–36

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. *Science of the Total Environment* 313: 77–89.

Manchanayake, P. and Madduma Bandara, C.M. (1999) *Water resources of Sri Lanka*. National Science Foundation, Colombo. Sri Lanka.

Oketola, A. A., Adekolurejo, S. M. and Osibanjo, O. (2013) Water Quality Assessment of River Ogun Using Multivariate Statistical Techniques. *Journal of Environmental Protection* 4: 466-479.

Thareja, S. and Trivedi, P. (2010) Assessment of Water Quality of Bennithora River in Karnataka through Multivariate Analysis. *Nature and Science* 8(6): 51 – 56.

Wang, G., Yinglan, A., Xu, Z. and Zhang, S. (2014) The influence of land use patterns on water quality at multiple spatial scales in a river system. *Hydrological processes* 28: 5259-5272.

Yang, LI., Linyu, X. U. and Shun, L. I. (2009) Water Quality Analysis of the Songhua River Basin Using Multivariate Techniques. *Journal of Water Resource and Protection* 2: 110-121.