Assessment of surface water quality using environmetric techniques: A case study of the Lower Songkhram river basin, Thailand

Sangam Shrestha[1] Somphinith Muangthong[2]

[1]Water Engineering and Management, Asian Institute of Technology, Thailand
[2]Rajamangala University of Technology Isan, Thailand

Abstract: Environmetric techniques such as cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) were applied for the assessment of spatial and temporal variations of a large complex water quality data set of the Songkhram River Basin, generated during 15 years (1995–2009) by monitoring of 17 parameters at 5 different sites. Hierarchical CA grouped 5 sampling sites into three clusters, i.e., upper stream (US), middle stream (MS) and lower stream (LS) sites, based on water quality characteristics. FA/PCA applied to the data sets thus the **obtained resulted in six latent factors explaining 80.80, 73.95 and 73.78% of the total variance in water quality data sets of LS, MS and US areas, respectively. This study highlights the usefulness of multivariate statistical assessment of complex databases in the identification of pollution sources and to better comprehend the spatial and temporal variations for effective river water quality management.**

Keywords: Factor analysis; Principal Component Analysis; Discriminant Analysis; Songkhram River Basin; Thailand

1.BACKGROUND AND METHODOLOGY

Application of environmetric techniques such as Cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) helps in the interpretation of complex data matrices to better understand the water quality and ecological status of the studied systems. It also allows for identification of possible factors/sources that influence water systems and offers a valuable tool for reliable management of water resources, both quantity and quality (Reghunath et al., 2002; Simeonova et al., 2003; Shrestha and Kazama, 2007; Shrestha et al., 2008).

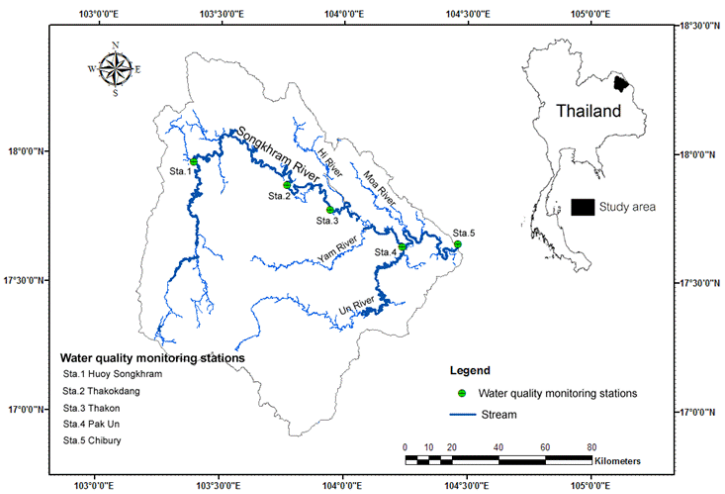

Fig .1. Location map of study area and water quality monitoring stations in the Songkhram River

In the present study, a large data matrix, obtained during a 15–year (1995–2009) monitoring program, is subjected to different environmetric techniques such as Cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) to extract information about the similarities and dissimilarities between sampling sites, identification of water quality variables responsible for spatial and temporal variations in river water quality, the hidden factors explaining the structure of the database, and the influence of possible sources (natural and anthropogenic) on the water quality parameters of the Songkhram River, Thailand (Fig. 1). The data sets of five water quality monitoring stations, comprising 17 water quality parameters monitored during wet–dry seasons over 15 years (1995–2009), were obtained from the Pollution Control Department, Thailand.

2. SPATIAL SIMILARITY AND SITE GROUPING ACCORDING TO WATER QUALITY

Cluster analysis (CA) was used to detect the similarity groups between the sampling sites. It yielded a dendrogram (Fig. 2), grouping all five sampling sites of the basin into three statistically significant clusters at (Dlink/Dmax) x 100 < 60. Since we used hierarchical agglomerative CA, the number of clusters was also decided by practicality of the results as there is ample information (e.g.

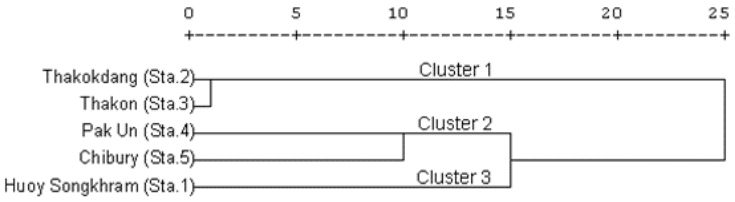

Fig. 2. Dendrogram showing clustering of sampling sites according to water quality characteristics of the Songkhram River

landuse) available on the study sites. Cluster 1 (Thakokdang (2) and Thakon (3)) corresponds to Middle stream (MS) and these stations receive pollution mostly from non-point sources, i.e., mostly from agricultural and orchard plantation activities. Cluster 2 (Pak Un (4) and Chiburi (5)) corresponds to Lower Stream (LS). These stations receive pollution from domestic wastewater and industrial effluents located in city areas (e.g. salt industrial unit). Cluster 3 (Houy Songkhram (1)) corresponds to relatively Upper stream (US). In cluster 3, one station, Houy Songkhram, is situated in the upstream site of the river. Results indicate that the CA technique is useful in offering reliable classification of surface waters in the whole region and will make it possible to design a future spatial sampling strategy in an optimal manner, which can reduce the number of sampling stations and associated costs.

3. TEMPORAL AND SPATIAL VARIATIONS IN RIVER WATER QUALITY

Temporal variations in water quality were evaluated through DA. Temporal DA was performed on raw data after dividing the whole data set into two seasonal groups (wet and dry).The temporal DA results suggest that hardness (CaCO3), nitrate nitrogen, total coliform bacteria, salinity, conductivity, turbidity, ammonia nitrogen, total phosphorus and water temperature are the most significant parameters to discriminate between the two seasons, which means that these nine parameters account for most of the expected temporal variations in the river water quality.

Spatial DA was performed with the same raw data set comprising 17 parameters after grouping into three major classes of LS, MS and US as obtained through CA. DA shows that turbidity, conductivity, salinity, biochemical oxygen demand, ammonia nitrogen, suspended solids and hardness ($CaCO_3$) are the discriminating parameters in space. Box and whisker plots of discriminating parameters identified by spatial DA (backward stepwise mode) were constructed to evaluate different patterns

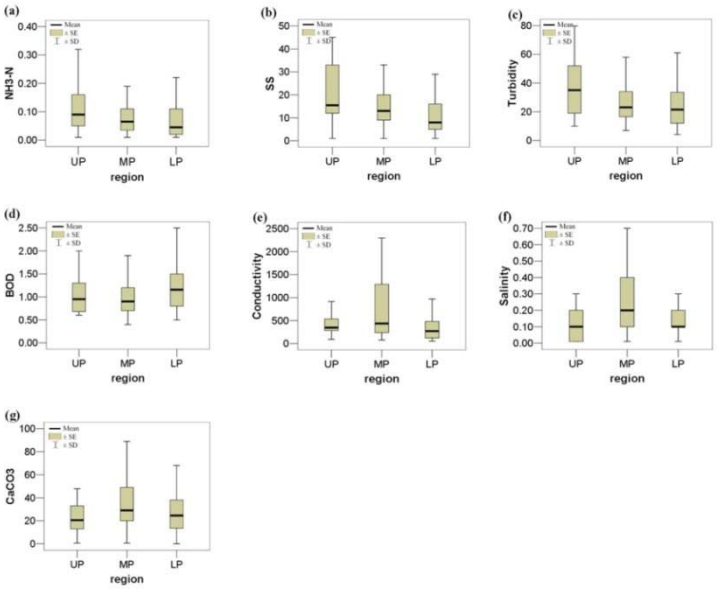

Fig. 3. Spatial variations: (a) NH_3-N, (b) SS, (c) turbidity, (d) BOD, (e) conductivity, (f) salinity, (g) $CaCO_3$ in water quality of the Songkhram River

associated with spatial variations in river water quality (Fig. 3). The trends for ammonia nitrogen (Fig. 3a), suspended solids (Fig. 3b) and turbidity (Fig. 3c) suggest a high load of dissolved organic matter from agricultural effluents located in the upper areas of the monitoring stations. This results in anaerobic conditions in the river, which in turn, results in formation of ammonia and organic acids. Hydrolysis of these

acidic materials causes a decrease of biochemical oxygen demand (Fig. 3d) in these sites. The conductivity (Fig. 3e), salinity (Fig. 3f) and hardness ($CaCO_3$) (Fig. 3g) are highest in MP, as they receive discharge from domestic wastewater and industrial effluents located in city areas. The average hardness ($CaCO_3$) of LS is higher than the US. However, the average conductivity and salinity of LS is lower than the US, which indicated the LS are geologically derived from the Maha Sarakham formation, comprised of a mixture of salt, shale and weathered sandstone.

4. DATA STRUCTURE DETERMINATION AND SOURCE IDENTIFICATION

PCA/FA was performed on the normalized data sets (17 variables) separately for the three different regions, viz., LS, MS and US, as delineated by CA techniques, to compare the compositional pattern between analyzed water samples and identify the factors influencing each one. The input data matrices (variables x cases) for PCA/FA were [17 x 60] for LS and MS and [17 x 30] for US. PCA of the three data sets yielded six PCs for the US, MS and LS with Eigenvalues >1, explaining 80.80, 73.95 and 73.78% of the total variance in respective water quality data sets. An Eigenvalue gives a measure of the significance of the factors: the factors with the highest Eigenvalues are the most significant. Eigenvalues of 1.0 or greater are considered significant (Kim and Mueller, 1987; Hair et al., 2006). Equal numbers of VFs were obtained for three sites through FA performed on the PCs. Corresponding VFs, variable loadings and explained variance are presented in Table 1. Liu et al. (2003) classified the factor loadings as 'strong', 'moderate' and 'weak', corresponding to absolute loading values of >0.75, 0.75–0.50 and 0.50–0.30, respectively.

Table 1. Loading of experimental variables (17) on significant principal components for (a) UP, (b) MP and (c) LS data sets

Parameters	VF1	VF2	VF3	VF4	VF5	VF6
Upper part (UP) (six significant principal components)						
WT	-0.205	-0.317	0.384	-0.148	-0.030	**0.731**
pH	0.627	-0.382	0.149	-0.124	-0.377	-0.273
Turbidity	0.355	**0.721**	0.309	0.344	-0.002	0.164
Conductivity	**0.752**	0.051	-0.149	0.306	-0.033	0.017
Salinity	0.074	0.001	-0.008	0.093	**0.938**	-0.040
DO	-0.043	**0.795**	0.046	-0.336	0.078	-0.134
BOD	**0.737**	-0.065	0.194	-0.203	0.119	0.068
TCB	0.151	0.151	**0.870**	-0.033	-0.034	-0.083
FCB	-0.024	0.054	**0.899**	0.221	0.093	0.085
TP	0.330	0.239	-0.357	0.083	0.072	0.689
NO_3-N	-0.347	0.619	0.127	0.545	0.082	-0.137
NO_2-N	0.177	0.657	0.131	0.086	0.125	-0.124
NH_3-N	0.349	0.147	0.437	**0.736**	0.099	0.029
TS	**0.916**	0.277	0.060	0.147	0.013	-0.004
TDS	**0.867**	0.199	0.022	0.140	-0.072	0.056
SS	0.099	**0.835**	-0.129	0.191	-0.207	0.245
$CaCO_3$ (Hardness)	0.384	-0.071	-0.243	0.508	-0.602	-0.203
Eigenvalue	3.831	3.142	2.323	1.651	1.501	1.288
% Total variance	22.537	18.483	13.664	9.709	8.831	7.577
Cumulative % variance	22.537	41.020	54.684	64.393	73.224	80.801
Weight	0.279	0.229	0.169	0.120	0.109	0.094
Middle part (MP) (six significant principal components)						
WT	0.033	-0.127	-0.085	-0.024	**0.802**	0.186
pH	0.531	0.135	0.099	-0.503	0.127	0.111
Turbidity	0.097	**0.831**	0.218	0.143	-0.179	-0.022
Conductivity	**0.858**	0.104	-0.125	0.078	0.101	0.070
Salinity	0.074	0.092	0.372	0.069	0.678	-0.184
DO	0.471	0.188	-0.419	-0.129	-0.083	0.401
BOD	0.055	0.362	-0.269	0.485	-0.046	0.455
TCB	-0.013	-0.037	**0.863**	-0.172	-0.024	0.089
FCB	-0.045	-0.143	**0.871**	0.099	0.087	-0.046
TP	-0.016	**0.722**	-0.300	0.291	0.045	-0.258
NO_3-N	-0.100	0.435	0.111	0.486	-0.530	0.033
NO_2-N	0.035	0.190	-0.019	0.102	-0.138	**-0.836**
NH_3-N	0.096	0.081	0.029	**0.828**	0.048	-0.095
TS	**0.962**	-0.010	-0.010	0.095	-0.025	-0.057
TDS	**0.952**	-0.100	-0.008	-0.054	0.091	-0.026
SS	0.000	**0.828**	-0.322	-0.177	0.003	0.029
$CaCO_3$ (Hardness)	0.463	-0.008	0.210	-0.001	-0.254	0.452
Eigenvalue	3.328	2.380	2.221	1.637	1.557	1.449
% Total variance	19.575	14.003	13.066	9.627	9.158	8.521
Cumulative % variance	19.575	33.577	46.643	56.270	65.428	73.949
Weight	0.265	0.189	0.177	0.130	0.124	0.115
Lower part (LS) (six significant principal components)						
WT	-0.091	-0.147	0.230	**0.758**	0.159	0.130
pH	0.534	-0.128	-0.028	0.080	-0.544	-0.245
Turbidity	0.153	**0.872**	0.247	-0.152	0.085	-0.032
Conductivity	**0.865**	0.043	-0.043	0.091	0.122	0.115
Salinity	0.151	-0.116	0.050	0.022	0.031	**0.789**
DO	0.405	0.145	-0.021	0.093	**0.709**	0.072
BOD	0.229	0.319	0.092	0.567	-0.383	0.232
TCB	0.008	-0.023	**0.888**	-0.032	-0.039	0.019
FCB	-0.096	-0.008	**0.900**	0.152	0.063	0.099
TP	-0.105	0.644	-0.345	-0.067	-0.076	0.131
NO_3-N	-0.181	0.514	0.236	-0.344	0.458	-0.305
NO_2-N	-0.007	0.324	0.170	-0.692	0.064	0.286
NH_3-N	0.197	0.532	0.408	-0.063	-0.278	-0.242
TS	**0.900**	0.152	0.006	-0.077	0.062	0.088
TDS	**0.872**	-0.050	0.024	-0.008	-0.018	0.058
SS	0.000	**0.805**	-0.177	-0.033	0.305	-0.128
$CaCO_3$ (Hardness)	0.640	-0.039	-0.056	-0.011	-0.138	-0.549
Eigenvalue	3.377	2.680	2.131	1.581	1.412	1.362
% Total variance	19.864	15.765	12.536	9.302	8.307	8.010
Cumulative % variance	19.864	35.629	48.166	57.468	65.774	73.784
Weight	0.269	0.214	0.170	0.126	0.113	0.109

Bold values indicate strong and moderate loadings.

For the data set pertaining to water quality in US, among the six VFs, VF1, explaining 22.54% of total variance, has strong positive loadings on conductivity, total solids and total dissolved solids and moderate positive loading on biochemical oxygen demand. VF2, explaining 18.48% of total variance, has strong positive loadings on dissolved oxygen

and suspended solid and moderate positive loadings on turbidity. VF3, explaining about 13.66% of total variance, has strong positive loadings on total coliform bacteria and fecal coliform bacteria. VF4, explaining about 9.71% of total variance, has moderate positive loading on ammonia nitrogen. This factor represents the contribution of non-point source pollution from forested areas. VF5, explaining 8.83% of total variance, has strong positive loadings on salinity. VF6 (7.58%) has moderate positive loadings on water temperature. This factor represents the seasonal effect of temperature.
For the data set representing the MS, among the total six significant VFs, VF1, explaining about 19.58% of total variance, has strong positive loadings on conductivity, total solid and total dissolved solid. VF2, explaining 14.00% of the total variance, has strong positive loadings on turbidity and suspended solid and moderate positive loading on total phosphorus. This factor represents the erosion effect during cultivation of soil and total phosphorus. VF3, explaining about 13.07% of total variance, has strong positive loadings on total coliform bacteria and fecal coliform bacteria. VF4, explaining about 9.62% of total variance, has strong positive loading on ammonical nitrogen. This factor represents the contribution of non–point source pollution from orchard and agricultural areas. In these areas, farmers use the nitrogenous fertilizer, which undergo nitrification processes, and the rivers receive nitrate nitrogen via groundwater leaching. VF5, explaining 9.16% of total variance, has strong positive loadings on water temperature. VF6 (8.52%) has strong negative loadings on nitrite–nitrogen.

Lastly, for the data set pertaining to LS, among the six VFs, VF1, explaining 19.86% of total variance, has strong positive loading on total solid, total dissolved solid and conductivity. VF2, explaining 15.77% of the total variance, has strong positive loadings on turbidity and suspended solid. VF1 and VF2 represent the seasonal impact of turbidity and conductivity. This factor explains the erosion from upland areas during rainfall events and the positive correlation with turbidity and conductivity indicates the loading of partially decayed organic matters from agricultural areas. VF3, explaining 12.54% of the total variance, has strong positive loadings for total coliform bacteria and fecal coliform bacteria. VF4, explaining 9.30% of total variance, has strong positive loading on water temperature, which represents the seasonal impact of temperature. VF5, explaining 8.31% of total variance, has moderate positive loading on dissolved oxygen. VF6, explaining the lowest variance (8.01%), has strong positive loadings on salinity which represents the physiochemical source of variability.

5. CONCLUSIONS

In this case study, environmetric techniques were used to evaluate spatial and temporal variations in surface water quality of the Songkhram River. Hierarchical CA grouped five sampling sites into three clusters of similar water quality characteristics. Based on obtained information, it is possible to design optimal sampling strategies which could reduce the number of sampling stations and associated costs. Although the FA/PCA did not result in a significant data reduction, it helped extract and identify the factors/sources responsible for variations in river water quality at three different sampling sites. VFs obtained from factor analysis indicated that the parameters responsible for water quality variations are mainly related to temperature (natural), organic pollution (point source: domestic wastewater) in relatively US, organic pollution (point source: domestic wastewater) and nutrients (non-point sources: agriculture and orchard plantations) in MS, and organic pollution and nutrients (point sources: domestic wastewater) in LS on the basin.

REFERENCES:

[1] Reghunath, R., Murthy, T.R.S., & Raghavan, B.R. (2002). The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. Water Research, 36, 2437-2442.

[2] Simeonov, V., Stratis J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., & Sofoniou, M. (2003). Assessment of the surface water quality in Northern Greece. Water Research, 37, 4119-4124.

[3] Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environmental Modelling & Software, 22, 464–475.

[4] Shrestha, S., Kazama, F., & Nakamura, T. (2008). Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River. Journal of Hydroinformatics, 10, 43–54.