

Assessment of surface water quality using multivariate statistical techniques: case study of the Nampong River and Songkhram River, Thailand

Somphinit Muangthong · Sangam Shrestha

Received: 4 December 2013 / Accepted: 22 July 2015
© Springer International Publishing Switzerland 2015

Abstract Multivariate statistical 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 Nampong River and Songkhram River, generated for more than 10 years (1996–2012) by monitoring of 16 parameters at different sites. According to the water quality characteristics, hierarchical CA grouped 13 sampling sites of the Nampong River into two clusters, i.e., upper stream (US) and lower stream (LS) sites, and five sampling sites of the Songkhram River into three clusters, i.e., upper stream (US), middle stream (MS) and lower stream (LS) sites. PCA/FA applied to the data sets thus obtained five latent factors explaining 69.80 and 69.32 % of the total variance in water quality data sets of LS and US areas, respectively, in the Nampong River and 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, in the Songkhram River. This study highlights the

usefulness of multivariate statistical assessment of complex databases in the identification of pollution sources to better comprehend the spatial and temporal variations for effective river water quality management.

Keywords Surface water quality · Factor analysis · Principal component analysis · Discriminant analysis · Nampong River · Songkhram River · Thailand

Introduction

Water quality is influenced by natural processes including precipitation rate, weathering processes, and sediment transport and also by anthropogenic activities including urban development and expansion, and industrial and agricultural practices. Pollution of surface water with toxic chemicals and excess nutrients, resulting from storm water runoff, vadose zone leaching, and groundwater discharges, has been an issue of worldwide environmental concern. Therefore, effective and long-term management of surface water requires a fundamental understanding of hydro-morphological, chemical, and biological characteristics. However, due to spatial and temporal variations in water quality, a monitoring program, providing a representative and reliable estimation of the quality of surface waters, is necessary (Shrestha and Kazama 2007; Guangjia et al. 2010; Muangthong 2015).

Application of multivariate statistical techniques such as cluster analysis (CA), principal component analysis (PCA), factor analysis (FA), and discriminant analysis (DA) has increased tremendously in recent years for

S. Muangthong
Faculty of Engineering and Architecture, Rajamangala University of Technology Isan, 744, Naimoung, Moung, Nakornratchasima 30000, Thailand

S. Shrestha (✉)
Water Engineering and Management, School of Engineering and Technology, Asian Institute of Technology, PO Box 4, Khlong Luang, Pathumthani 12120, Thailand
e-mail: sangam@ait.asia

S. Shrestha
e-mail: sangamshrestha@gmail.com

analyzing environmental data and drawing meaningful information (Reghunath et al. 2002; Zhou et al. 2007; Boyacioglu and Boyacioglu 2007; Coletti et al. 2010; Shihab and Abdul Baqi 2010; Guangjia et al. 2010; Batayneh and Zumlot 2012; Shrestha and Muangthong 2014). Application of different multivariate statistical techniques 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; Simeonov et al. 2003; Shrestha and Kazama 2007; Shrestha et al. 2008; Alexakis 2011; Batayneh and Zumlot 2012).

In the present study, a large data matrix, obtained for more than 10 years (1996–2012) of monitoring program, is subjected to different multivariate statistical techniques 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 Nampong River and Songkhram River in Thailand.

Materials and methods

Study area

The Nampong River Basin and Songkhram River Basin are located in the northeast region of Thailand (Fig. 1). The Nampong River is 220 km long and drains the basin area of 12,560 km². The basin is relatively flat with an average elevation of 300 m above mean sea level (masl) while the western edge is rougher and rises to heights of 1300 masl. The upper Nampong River Basin yields water for the Ubolratana reservoir that provides downstream areas with hydropower as well as irrigated water for agriculture. Most underlying rocks in the basin are of sandstone or limestone. Surface soils are mainly sandy loam with sandy clay sub-soils and contain more clay in the west and more sand in the east. Stream flow declines very rapidly with the end of the rainy season, indicating poor soil moisture retention. Underlying geology in the basin consists mainly of limestone and sandstone which are notorious for deep leakage.

The Songkhram River is 420 km in length and drains the basin area of 13,081 km². It rises in the western Phu Phan mountain range at an altitude of around 400 masl and flows through parts of Udon Thani, Nong Khai, Sakon Nakhon, and Nakhon Phanom provinces. The lower part of the Songkhram River meanders over an extensive floodplain at an altitude of 145–160 masl with a gentle channel gradient of about 1:30,000 (Blake and Pitakthepsombut 2006a, b). The Songkhram River meets with the Mekong River and is in the Tha Utaen district of Nakhon Phanom province, about 40 km north of Nakhon Phanom city. Several major tributaries join the Songkhram River from the north (e.g., Hi and Mao rivers) and from the south (Un and Yam rivers) forming one extensive lowland floodplain system.

The Nampong River Basin and Songkhram River Basin experience a tropical, semi-arid climate, with three distinct seasons. There are six to seven rainy season months (May–October). Average annual rainfall within the basin varies considerably, with southern parts of the basin receiving less than 1200 mm per annum, rising to over 2100 mm per annum in the far northern part of the basin (Sombutputorn 1998). The rainy season normally peaks in August to September, when floods reach their maximum extent. The cool season extends from November to February and is marked by generally dry and cool air from the northeast monsoon. Minimum temperatures rarely fall below 12 °C in the cool season. The hot season extends from March to mid or late May, if the rains arrive late. The early part tends to be very dry and warm, marked by occasional thunderstorms, and as maximum daytime temperatures climb to over 40 °C by mid to late April, the intensity and frequency of thunder showers increases. Annual evaporation rate ranges from 100 to 165 mm.

The Pollution Control Department (PCD) of Thailand has been collecting various water quality parameters from five water quality monitoring stations in the Songkhram River and 13 water quality monitoring stations in the Nampong River. The stations were selected from the river water quality monitoring network, which covers a wide range of catchments and surface water types (rivers, streams, and tributaries).

Monitored parameters and analytical methods

In this study, the data sets of 13 water quality monitoring stations of the Nampong River and five water quality monitoring stations of the Songkhram River comprising

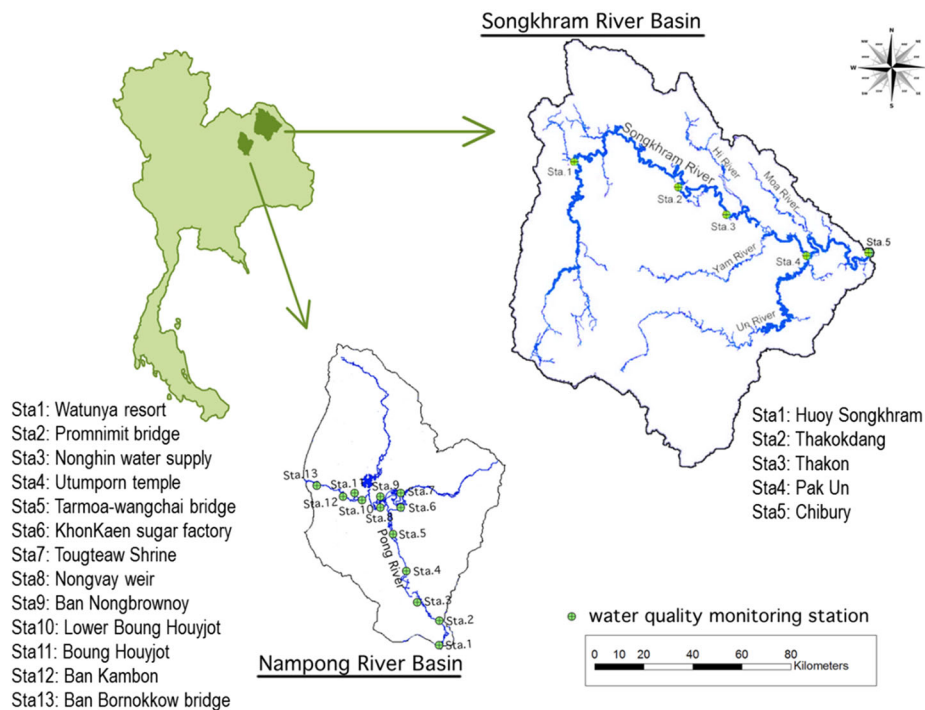


Fig. 1 Location map of study area and water quality monitoring stations in the Nampong River and Songkhram River (the Nampong River consists of 13 stations and the Songkhram River consists of 5 stations)

16 water quality parameters monitored monthly interval during wet–dry seasons were obtained from the PCD. The 16 parameters out of 29 were selected based on continuous data availability. The selected parameters include water temperature, pH, turbidity, conductivity, salinity, dissolved oxygen, 5-day biochemical oxygen demand, total coliform bacteria, fecal coliform bacteria, total phosphorus, nitrate nitrogen, nitrite nitrogen, ammonia nitrogen, total solid, total dissolved solid, and suspended solid. The water quality parameters, their units, and methods of analysis are summarized in Table 1. The PCD has sampled, preserved, and analyzed all the water quality parameters according to the standards set by the US Environmental Protection Agency (USEPA).

Data and multivariate statistical methods

The Kolmogorov–Smirnov (K–S) statistics were used to test the goodness-of-fit of the data to log-normal distribution. According to the K–S test, all the variables are log-normally distributed with 95 % or higher confidence. Similarly, to examine suitability of the data for

PCA/FA, Kaiser–Meyer–Olkin (KMO) and Bartlett’s test were performed. KMO is a measure of sampling adequacy that indicates the proportion of variance, i.e., which might be caused by underlying factors. High value (close to 1) generally indicates that PCA/FA may be useful, which in this study KMO is 0.69 for the Nampong River and 0.87 for the Songkhram River. Bartlett’s test of sphericity indicates whether correlation matrix is an identity matrix, which would indicate that variables are unrelated. The significance level which is 0 in this study (less than 0.05) indicates that there are significant relationships among variables.

Spearman rank–order correlations (Spearman’s R coefficient) were used to study the correlation structure between variables to account for non-normal distribution of water quality parameters (Shrestha and Kazama 2007). In this study, temporal variations of river water quality parameters were first evaluated through a season parameter correlation matrix, using Spearman non-parametric correlation coefficients. The water quality parameters were grouped into two seasons: wet (May–October) and dry (November–April), and each assigned a numerical value in the data file (wet = 1 and dry = 2),

Table 1 Water quality parameters, units, and analytical methods used for water of the Nampong River and Songkhram River

S.N.	Parameters	Abbreviations	Units	Analytical methods
1	Water temperature	WT	°C	Thermometer
2	pH	pH	pH	Electrometric (pH meter)
3	Turbidity	Tur	NTU	Turbidity meter
4	Conductivity	Cond	μS/cm	Electrometric (conductivity meter)
5	Salinity	Sal	ppt	Electrometric (conductivity meter)
6	Dissolved oxygen	DO	mg/l	Azide modification
7	Biochemical oxygen demand	BOD	mg/l	Azide modification at 20 °C (5 days)
8	Total coliform bacteria	TCB	MPN/100 ml	Multiple tube fermentation technique
9	Fecal coliform bacteria	FCB	MPN/100 ml	Multiple tube fermentation technique
10	Total phosphorus	TP	mg/l	Ascorbic acid
11	Nitrate nitrogen	NO ₃ -N	mg/l	Cadmium reduction
12	Nitrite nitrogen	NO ₂ -N	mg/l	Distillation nesslerization
13	Ammonia nitrogen	NH ₃ -N	mg/l	Distillation nesslerization
14	Total solid	TS	mg/l	Total residue dried at 103–105 °C
15	Total dissolved solid	TDS	mg/l	Total dissolved solids dried at 103–105 °C
16	Suspended solid	SS	mg/l	Suspended solids dried at 103–105 °C

which, as a variable corresponding to the season, was correlated (pair by pair) with all the measured parameters.

River water quality data sets were subjected to four multivariate techniques: CA, PCA, FA, and DA (Reghunath et al. 2002; Zhou et al. 2007; Boyacioglu and Boyacioglu 2007; Shrestha et al. 2008; Zhang et al. 2011; Shrestha and Muangthong 2014; Muangthong 2015). DA was applied to raw data, whereas PCA, FA, and CA were applied to experimental data, standardized through *z* scale transformation to avoid misclassifications arising from the different orders of magnitude of both numerical values and variance of the parameters analyzed (Liu et al. 2003; Simeonov et al. 2003).

Cluster analysis

Cluster analysis (CA) is a group of multivariate techniques whose primary purpose is to assemble objects with respect to a predetermined selection criterion resulting in high internal (within-cluster) homogeneity and high external (between clusters) heterogeneity. Hierarchical agglomerative clustering is the most common approach, which provides intuitive similarity relationships between any one sample and the entire data set, and is typically illustrated by a dendrogram (tree diagram) (McKenna 2003). The Euclidean distance usually

gives the similarity between two samples and a distance can be represented by the difference between analytical values from the samples (Shrestha et al. 2008). In this study, hierarchical agglomerative CA was performed on the normalized data set by means of the Ward's method, using squared Euclidean distances as a measure of similarity. The Ward's method uses an analysis of variance approach to evaluate the distances between clusters in an attempt to minimize the sum of squares (SS) of any two clusters that can be formed at each step. The spatial variability of water quality in the whole river basin was determined from CA, using the linkage distance, reported as D_{link}/D_{max} , which represents the quotient between the linkage distances for a particular case divided by the maximal linkage distance. The quotient is then multiplied by 100 as a way to standardize the linkage distance represented on the *y*-axis (Simeonov et al. 2003; Singh et al. 2005; Shrestha and Kazama 2007).

Principal component analysis/factor analysis

Principal component analysis (PCA) is designed to transform the original variables into new, uncorrelated variables (axes), called the principal components (PCs), which are linear combinations of the original variables. PCA provides an objective way of finding indices of this type so that the variation in the data can be accounted for

as concisely as possible (Sarbu and Pop 2005). PC provides information on the most meaningful parameters, which describe a whole data set affording data reduction with minimum loss of original information (Helena et al. 2000).

Factor analysis (FA) follows PCA. The main purpose of FA is to reduce the contribution of less significant variables to simplify even more of the data structure coming from PCA by rotating the axis defined by PCA, according to well-established rules, and constructing new variables, also called varifactors (VF). PC is a linear combination of observable water quality variables, whereas VF can include unobservable, hypothetical, latent variables (Helena et al. 2000). PCA of the normalized variables was performed to extract significant PCs and to further reduce the contribution of variables with minor significance; these PCs were subjected to varimax rotation (raw) generating VFs (Singh et al. 2004, 2005; Love et al. 2004; Abdul-Wahab et al. 2005). As a result, a small number of factors will usually account for approximately the same amount of information as do the much larger set of original observations.

Discriminant analysis

Discriminant analysis (DA) is used to classify cases into categorical-dependent values, usually a dichotomy. If DA is effective for a set of data, the classification table of correct and incorrect estimates will yield a high correct percentage. In DA, multiple quantitative attributes are used to discriminate between two or more naturally occurring groups. In contrast to CA, DA provides statistical classification of samples and it is performed with prior knowledge of membership of objects to a particular group or cluster. Furthermore, DA helps in grouping samples sharing common properties.

In this study, two groups for temporal (two seasons) and three groups for spatial (three sampling regions) evaluations have been selected and the number of analytical parameters used to assign a measure from a monitoring site into a group (season or monitoring area). DA was performed on each raw data matrix using standard, forward stepwise, and backward stepwise modes in constructing discriminant functions (DFs) to evaluate both the spatial and temporal variations in river water quality of the basin. The site (spatial) and the season (temporal) were the grouping (dependent) variables, whereas all the measured parameters constituted the independent variables.

Results and discussion

Spatial similarity and site grouping

Cluster analysis (CA) was used to detect the similarity groups between the sampling sites. It yielded a dendrogram, grouping 13 sampling sites of the Nampong River into two and five sampling sites of the Songkhram River into three statistically significant clusters at $(D_{link}/D_{max}) \times 100 < 60$ (Fig. 2). 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., land use) available on the study sites.

In the Nampong River, cluster 1 corresponds to lower stream (LS) sites. This cluster includes stations Watunya resort, Promnimit bridge, Nonghin water supply, Utumporn temple, Tarmoa-wangchai bridge, and Khon Kaen sugar factory. These stations receive pollution from domestic wastewater and industrial effluents located in urban areas. Cluster 2 includes stations Tougteaw Shrine, Nongvay weir, Ban Nongbrownoy, Lower Boung Houyjo, Upper Boung Houyjo, Ban Kambon, and Ban Bornokkow bridge and corresponds to relatively upper stream (US) sites. In cluster 2, seven stations are situated at the upper part of the river basin. These stations receive pollution from agricultural and farm effluents located in agriculture areas.

In the Songkhram River, cluster 1 includes stations Thakokdang and Thakon 3 and 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 includes stations Pak Un and Chiburi and 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 includes station Houy Songkhram and corresponds to relatively upper stream (US). 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. There are other reports (Simeonov et al. 2003; Zhang et al. 2011; Coletti et al. 2010) where similar approach has successfully been applied to water quality programs.

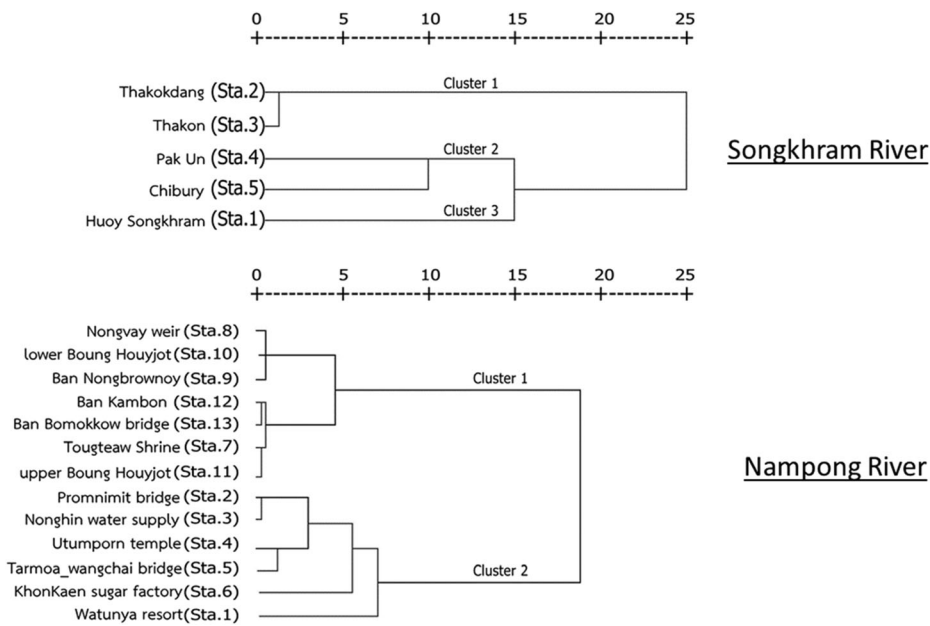


Fig. 2 Dendrogram showing clustering of sampling sites according to water quality characteristics of the Nampong River and Songkhram River

Temporal and spatial variations in river water quality

Temporal variations in river water quality parameters (Table 1) were evaluated through a season–parameter correlation matrix, which shows that most of the measured parameters were found to be significantly ($p < 0.01$) correlated with the season. Temporal variations in water quality were further evaluated through DA. Temporal DA was performed on raw data after dividing the whole data set into two seasonal groups (wet and dry). DFs and classification matrices (CMs) obtained from the standard, forward stepwise, and backward stepwise modes of DA are shown in Table 2. In forward stepwise mode, variables are included step-by-step beginning with the more significant until no significant changes are obtained, whereas, in backward stepwise mode, variables are removed step-by-step beginning with the less significant until no significant changes are obtained. Thus, the temporal DA results suggest that dissolved oxygen, biochemical oxygen demand, fecal coliform bacteria, nitrate nitrogen, nitrite nitrogen, suspended solid, turbidity, and conductivity are the most significant parameters to discriminate between the wet and dry seasons, which means that these eight parameters account for most of the expected temporal variations in the river water quality of the Nampong River. Similarly, nitrate nitrogen, total coliform bacteria, salinity,

conductivity, turbidity, ammonia nitrogen, total phosphorus, and water temperature are the most significant parameters to discriminate between the wet and dry seasons, which means that these nine parameters account for most of the expected temporal variations in the river water quality of the Songkhram River (Table 2).

The results of analysis on temporal variations in water quality can be used to design the water quality monitoring program. The discriminant analysis suggest that out of 16 parameters, only eight parameters (dissolved oxygen, biochemical oxygen demand, fecal coliform bacteria, nitrate nitrogen, nitrite nitrogen, suspended solid, turbidity, and conductivity) in the Nampong River and nine parameters (nitrate nitrogen, total coliform bacteria, salinity, conductivity, turbidity, ammonia nitrogen, total phosphorus, and water temperature) in the Songkhram River can be included in the future water quality monitoring program as they are the most significant parameters to discriminate between the two seasons. However, the selection of water quality parameters depends on the objectives of the monitoring program.

Spatial DA was performed with the same raw data set comprising 16 parameters after grouping into two major classes of lower stream (LS) and upper stream (US) in the Nampong River and three major classes of LS, MS,

Table 2 Classification function coefficients for discriminant analysis of temporal variations in water quality of the Nampong River and Songkhram River

Parameters	Wet season coefficient	Dry season coefficient
Nampong River		
DO	2.210	2.890
BOD	2.205	2.937
FCB	0.000	0.001
NO ₃ -N	0.091	7.580
NO ₂ -N	18.852	44.522
SS	-0.010	0.013
Turbidity	0.023	0.063
Conductivity	0.070	0.082
(Constant)	-13.124	-23.399
Songkhram River		
WT	6.789	6.474
Turbidity	0.125	0.062
Conductivity	-0.008	-0.004
Salinity	12.352	4.100
TCB	0.000	0.000
TP	15.191	36.499
NO ₃ -N	9.006	14.173
NH ₃ -N	-21.126	-5.585
(Constant)	-102.834	-97.979

and US in the Songkhram River as obtained through CA. The sites (clustered) were the grouping (dependent) variables, while all the measured parameters constituted the independent variables. DFs are shown in Table 3. DA shows that dissolved oxygen, biochemical oxygen demand, fecal coliform bacteria, total phosphorus, nitrate nitrogen, nitrite nitrogen, suspended solid, total dissolve solid, water temperature, conductivity, and salinity are the discriminating parameters in space in the Nampong River whereas turbidity, conductivity, salinity, biochemical oxygen demand, ammonia nitrogen, and suspended solid are the discriminating parameters in space in the Songkhram River.

Data structure determination and source identification

PCA/FA was performed on the normalized data sets (16 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.

In the Nampong River, PCA of the two data sets yielded five PCs for the LS and US with Eigenvalues >1, explaining 69.33 and 69.81 % of the total variance in respective water quality data sets. An Eigenvalue gives a measure of the significance of the factor: 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 4. 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.

For the data set pertaining to US, among the five VFs, VF1, explaining 19.98 % of total variance, has strong positive loadings on turbidity, suspended solid, and total solid. This factor represents the contribution of non-point source pollution from the forest and agriculture areas. VF2, explaining 15.21 % of total variance, has strong negative loadings on pH and DO. This factor represents the contribution of non-point source pollution and indicates the loading of partially decayed organic matters from agricultural areas. VF3, explaining about 12.88 % of total variance, has strong positive loadings on conductivity and salinity. VF4, explaining about 12.24 % of total variance, has strong positive loading on total coliform bacteria and fecal coliform bacteria. This factor represents the contribution of point source pollution from human activities. VF5, explaining 9.02 % of total variance, has strong positive loadings on total phosphorus and ammonia nitrogen. This factor represents the contribution of non-point source pollution from agricultural farm.

For the data set pertaining to water quality in LS, among five VFs, VF1, explaining 15.99 % of total variance, has strong positive loading on total solid and total dissolved solid. This factor represents the contribution of point source pollution from urban areas. VF2, explaining 14.61 % of the total variance, has strong positive loadings on total phosphorus and nitrate nitrogen. VF3, explaining 14.11 % of the total variance, has strong negative loadings for pH and nitrite nitrogen. VF4, explaining 12.83 % of total variance, has strong positive loadings on turbidity, BOD, and suspended solid. VF5, explaining 12.26 % of total variance, has strong positive loadings on total

Table 3 Classification function coefficients for discriminant analysis of spatial variations in water quality of the Nampong River and Songkhram River

Nampong River			
Parameters	Upper part coefficient		Lower part coefficient
DO	1.809		2.810
BOD	−2.041		−0.137
FCB	0.003		0.001
TP	0.912		−5.149
NO ₃ -N	−12.801		−47.513
NO ₂ -N	9.801		5.084
TD	−0.117		−0.080
TDS	0.068		0.144
Water temperature	10.109		8.688
Conductivity	0.011		−0.008
Salinity	7.222		−14.458
(Constant)	−166.561		−134.005
Songkhram River			
Parameters	Upper stream coefficient	Middle stream coefficient	Lower stream coefficient
Turbidity	−0.006	−0.030	0.008
Conductivity	0.000	0.000	−0.001
Salinity	14.744	22.824	13.489
BOD	4.426	4.647	5.963
NH ₃ -N	21.544	18.132	9.916
SS	0.225	0.214	0.093
(Constant)	−8.712	−10.684	−7.622

coliform bacteria and fecal coliform bacteria. This factor represents the contribution of point source pollution such as toilet and kitchen water from urban areas.

In the Songkhram River, 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 (Table 5). 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 solid, and total dissolved solid and moderate positive loading on biochemical oxygen demand. This factor represents the contribution of non-point source pollution from the forest areas. VF2, explaining 18.48 % of total variance, has strong positive loadings on dissolved oxygen and suspended solid and moderate positive loadings on turbidity. This factor also represents the contribution from forest areas such as from erosion. VF3, explaining about 13.66 % of total variance, has strong positive loadings on total coliform bacteria and

fecal coliform bacteria. This factor represents the contribution of point source pollution such as toilet and kitchen water from urban areas. 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

Table 4 Loading of experimental variables (16) on significant principal components for (a) US, and (b) LS data sets of Nampong River

Parameters	VF1	VF2	VF3	VF4	VF5
Upper stream (US) (five significant principal components)					
WT	-0.466	0.546	0.115	0.107	0.113
pH	-0.040	-0.777	0.027	0.059	-0.203
Turbidity	0.722	-0.028	-0.455	-0.102	0.150
Conductivity	-0.019	0.339	0.756	0.040	0.223
Salinity	-0.185	-0.190	0.854	0.048	0.105
DO	-0.106	-0.842	-0.045	-0.055	-0.055
BOD	0.619	-0.042	0.022	-0.209	-0.053
TCB	-0.103	0.022	0.046	0.881	-0.004
FCB	-0.140	-0.042	-0.012	0.840	0.070
TP	0.148	0.400	0.093	0.265	0.669
NO ₃ -N	-0.074	0.486	0.483	0.087	0.085
NO ₂ -N	-0.137	0.448	0.171	0.558	0.113
NH ₃ -N	-0.045	0.121	0.182	-0.028	0.878
SS	0.754	-0.135	-0.330	0.001	0.095
TD	0.945	0.088	-0.096	-0.038	0.103
TDS	0.699	0.173	0.334	-0.099	-0.190
Eigenvalue	4.058	2.966	1.718	1.339	1.012
% total variance	19.983	15.209	12.877	12.239	9.019
Cumulative % variance	19.983	35.193	48.070	60.309	69.327
Lower stream (LS) (five significant principal components)					
WT	-0.490	0.362	0.475	-0.194	0.080
pH	0.094	0.068	-0.805	-0.012	-0.061
Turbidity	0.092	-0.011	-0.072	0.717	0.020
Conductivity	0.188	0.609	0.439	-0.092	-0.100
Salinity	-0.009	-0.621	0.082	-0.127	0.178
DO	0.082	0.109	-0.453	0.505	0.199
BOD	0.187	-0.156	-0.008	0.696	-0.073
TCB	-0.132	0.081	0.088	-0.038	0.882
FCB	0.019	0.113	0.027	-0.033	0.898
TP	-0.009	0.697	0.292	-0.016	0.060
NO ₃ -N	-0.268	0.685	-0.058	0.059	0.091
NO ₂ -N	0.065	0.322	-0.727	0.028	0.071
NH ₃ -N	-0.104	0.625	-0.149	-0.209	0.324
SS	-0.005	-0.024	0.076	0.702	-0.088
TD	0.942	-0.013	-0.013	0.173	-0.051
TDS	0.955	-0.072	-0.011	0.052	-0.047
Eigenvalue	3.495	2.031	1.903	1.566	1.105
% total variance	15.998	14.610	14.111	12.829	12.258
Cumulative % variance	15.998	30.608	44.719	57.548	69.806

Italic values indicate strong and moderate loadings

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.

Table 5 Loading of experimental variables (16) on significant principal components for (a) US, (b), MS and (c) LS data sets

Parameters	VF1	VF2	VF3	VF4	VF5	VF6
Upper stream (US) (six significant principal components)						
WT	-0.205	-0.317	0.384	-0.148	-0.030	<i>0.731</i>
pH	0.627	-0.382	0.149	-0.124	-0.377	-0.273
Turbidity	0.355	<i>0.721</i>	0.309	0.344	-0.002	0.164
Conductivity	<i>0.752</i>	0.051	-0.149	0.306	-0.033	0.017
Salinity	0.074	0.001	-0.008	0.093	<i>0.938</i>	-0.040
DO	-0.043	<i>0.795</i>	0.046	-0.336	0.078	-0.134
BOD	<i>0.737</i>	-0.065	0.194	-0.203	0.119	0.068
TCB	0.151	0.151	<i>0.870</i>	-0.033	-0.034	-0.083
FCB	-0.024	0.054	<i>0.899</i>	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	<i>0.736</i>	0.099	0.029
TS	<i>0.916</i>	0.277	0.060	0.147	0.013	-0.004
TDS	<i>0.867</i>	0.199	0.022	0.140	-0.072	0.056
SS	0.099	<i>0.835</i>	-0.129	0.191	-0.207	0.245
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
Middle stream (MS) (six significant principal components)						
WT	0.033	-0.127	-0.085	-0.024	<i>0.802</i>	0.186
pH	0.531	0.135	0.099	-0.503	0.127	0.111
Turbidity	0.097	<i>0.831</i>	0.218	0.143	-0.179	-0.022
Conductivity	<i>0.858</i>	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	<i>0.863</i>	-0.172	-0.024	0.089
FCB	-0.045	-0.143	<i>0.871</i>	0.099	0.087	-0.046
TP	-0.016	<i>0.722</i>	-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	<i>0.828</i>	0.048	-0.095
TS	<i>0.962</i>	-0.010	-0.010	0.095	-0.025	-0.057
TDS	<i>0.952</i>	-0.100	-0.008	-0.054	0.091	-0.026
SS	0.000	<i>0.828</i>	-0.322	-0.177	0.003	0.029
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
Lower stream (LS) (six significant principal components)						
WT	-0.091	-0.147	0.230	<i>0.758</i>	0.159	0.130
pH	0.534	-0.128	-0.028	0.080	-0.544	-0.245
Turbidity	0.153	<i>0.872</i>	0.247	-0.152	0.085	-0.032

Table 5 (continued)

Parameters	VF1	VF2	VF3	VF4	VF5	VF6
Conductivity	<i>0.865</i>	0.043	−0.043	0.091	0.122	0.115
Salinity	0.151	−0.116	0.050	0.022	0.031	<i>0.789</i>
DO	0.405	0.145	−0.021	0.093	<i>0.709</i>	0.072
BOD	0.229	0.319	0.092	0.567	−0.383	0.232
TCB	0.008	−0.023	<i>0.888</i>	−0.032	−0.039	0.019
FCB	−0.096	−0.008	<i>0.900</i>	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	<i>0.900</i>	0.152	0.006	−0.077	0.062	0.088
TDS	<i>0.872</i>	−0.050	0.024	−0.008	−0.018	0.058
SS	0.000	<i>0.805</i>	−0.177	−0.033	0.305	−0.128
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

Italic values indicate strong and moderate loadings

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.

Conclusions

In this case study, different multivariate statistical techniques were used to evaluate spatial and temporal variations in surface water quality of the Nampong River and Songkhram River. Hierarchical CA grouped 13 sampling sites into two clusters and five sampling sites into three clusters of similar water quality characteristics in the Nampong River and Songkhram River, respectively. Based on obtained information, it is possible to design optimal sampling strategy which could reduce the number of sampling stations and associated costs. Although the PCA/FA did not result in a significant data reduction, it helped extract and identify the factors/sources responsible for variations in river water quality at different sampling sites. Varifactors obtained from factor analysis indicate that the parameters responsible for water quality variations are mainly related to organic pollution and nutrients in upper stream, and organic pollution and nutrients in lower part areas in the Nampong River whereas 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 Songkhram River.

Acknowledgments The authors wish to thank Faculty of Engineering and Architecture, Rajamangala University of Technology Isan for providing financial assistance. The authors also sincerely thank Pollution Control Department (PCD), Thailand, for providing river water quality monitoring data.

References

- Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10), 1263–1271.
- Alexakis, D. (2011). Assessment of water quality in the Messolonghi–Etoliko and Neochorio region (West Greece) using hydrochemical and statistical analysis methods. *Environmental Monitoring and Assessment*, 182(1–4), 397–413.
- Blake, D., & Pitakthepsombut, R. (2006a). Situation analysis: lower Songkhram river basin, Thailand: a publication of the Mekong Wetlands biodiversity conservation and sustainable use programme.
- Batayneh, A., & Zumlot, T. (2012). Multivariate statistical approach to geochemical methods in water quality factor identification; application to the shallow aquifer system of the Yarmouk basin of north Jordan. *Research Journal of Environmental and Earth Sciences*, 4(7), 756–768.
- Blake, D., & Pitakthepsombut, R. (2006b). Situation analysis: lower Songkhram river basin, Thailand: a publication of the Mekong wetlands biodiversity conservation and sustainable use programme.
- Boyacioglu, H., & Boyacioglu, H. (2007). Water pollution sources assessment by multivariate statistical methods in the Tahtali basin, Turkey. *Environmental Geology*, 54, 275–282.
- Coletti, C., Testezlaf, R., Ribeiro, T. A. P., Souza, R. T. G., & Pereira, D. A. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 14, 517–522.
- Hair, J. E., William, C. B., Barry, J. B., Rolph, E. A., & Tatham, R. L. (2006). *Multivariate data analysis* (6 ed.,). New Jersey: Pearson Printice Hall.
- Helena, B., Pardo, R., Vega M., Barrado, E., Fernandez, J.M., & Fernandez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Research*, 34, 807–816.
- Guangjia, J., Dianwei, L., Kaishan, S., Zongming, W., Bai, Z., & Yuandong, W. (2010). Application of multivariate model based on three simulated sensors for water quality variables estimation in Shitoukoumen reservoir, Jilin province, China. *Chinese Geographical Science*, 20(4), 337–344.
- Kim, J.-O., & Mueller, C. W. (1987). *Introduction to factor analysis: what it is and how to do it. Quantitative applications in the social sciences series*. Newbury Park: Sage University Press.
- Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. *Science in the Total Environment*, 313, 77–89.
- Love, D., Hallbauer, D., Amos, A., & Hranova, R. (2004). Factor analysis as a tool in groundwater quality management: two southern African case studies. *Physics and Chemistry of the Earth*, 29, 1135–1143.
- McKenna Jr., J. E. (2003). An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. *Environmental Modelling & Software*, 18(3), 205–220.
- Muangthong, S. (2015). Assessment of surface water quality using multivariate statistical techniques: a case study of the Nampong river basin, Thailand. *The Journal of Industrial Technology*, 11(1), 25–37.
- 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.
- Sarbu, C., & Pop, H. F. (2005). Principal component analysis versus fuzzy principal component analysis. A case study: the quality of Danube water (1985–1996). *Talanta*, 65, 1215–1220.
- Shihab, A. S., & Abdul Baqi, Y. T. (2010a). Multivariate analysis of ground water quality of Makhmor plain/north Iraq. *Damascus University Journal*, 26(1), 19–26.
- Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsas, D., Anthemidis, A., & Sofoniou, M. (2003). Assessment of the surface water quality in northern Greece. *Water Research*, 37, 4119–4124.
- Singh, K. P., Amrita, M., Dinesh, M., & Sarita, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti river (India) a case study. *Water Research*, 38, 3980–3992.
- Singh, K. P., Malik, A., & Sinha, S. (2005). Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques: a case study. *Analytica Chimica Acta*, 538, 355–374.
- 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.
- 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.
- Shrestha, S., & Muangthong, S. (2014). Assessment of surface water quality of Songkhram river (Thailand) using environmetric techniques. *International Journal of River Basin Management*, 12(4), 341–356.
- Sombutputorn N. (1998). An application of remotely sensed data and geographic information system for wetland ecosystem mapping. Master thesis of Korn Kaen University.
- Zhang, X., Wang, Q., Liu, Y., Wu, J., & Yu, M. (2011). Application of multivariate statistical techniques in the assessment of water quality in the southwest new territories and Kowloon, Hong Kong. *Environmental Monitoring and Assessment*, 173(1–4), 17–27.
- Zhou, F., Liu, Y., & Guo, H. C. (2007). Application of multivariate statistical methods to the water quality assessment of the watercourses in the northwestern new territories, Hong Kong. *Environmental Monitoring and Assessment*, 132(1–3), 1–13.