Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River

Sangam Shrestha, Futaba Kazama and Takashi Nakamura

ABSTRACT

Multivariate statistical techniques, such as principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA), were applied for the evaluation of temporal/spatial variations and the interpretation of a large complex water quality dataset of the Mekong River using data sets generated during 6 years (1995–2000) of monitoring of 18 parameters (16,848 observations) at 13 different sites. The results of PCA/FA revealed that most of the variations are explained by dissolved mineral salts along the whole Mekong River and in individual stations. Discriminant analysis showed the best results for data reduction and pattern recognition during both spatial and temporal analysis. Spatial DA revealed 8 parameters (total suspended solids, calcium, sodium, alkalinity, chloride, iron, nitrate nitrogen, total phosphorus) and 12 parameters (total suspended solids, calcium, sodium, potassium, alkalinity, chloride, sulfate, iron, nitrate nitrogen, total phosphorus, silicon, dissolved oxygen) are responsible for significant variations between monitoring regions and countries, respectively. Temporal DA revealed 3 parameters (conductivity, alkalinity, nitrate nitrogen) between monitoring regions; 3 parameters (total suspended solids, conductivity, silicon) in midstream region; and 2 parameters (conductivity, silicon) in upstream, lower stream and delta region which are the most significant parameters to discriminate between the four different seasons (spring, summer, autumn, winter). Thus, this study illustrates the usefulness of principal component analysis, factor analysis and discriminant analysis for the analysis and interpretation of complex datasets and in water quality assessment, identification of pollution sources/factors, and understanding of temporal and spatial variations of water quality for effective river water quality management.

Key words | discriminant analysis, factor analysis, Mekong River, principal component analysis, transboundary river, water quality

Sangam Shrestha (corresponding author)
Futaba Kazama
Takashi Nakamura
Department of Ecosocial System Engineering,
Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi,
4-3-11, Takeda, Kofu,
Yamanashi 400-8511,
Japan
E-mail: *sangam@yamanashi.ac.jp*

INTRODUCTION

A river is a system comprising both the main course and the tributaries, carrying the one-way flow of a significant load of matter in dissolved and particulate phases from both natural and anthropogenic sources. The quality of a river at any point reflects several major influences, including the lithology of the basin, atmospheric inputs, climatic conditions and anthropogenic inputs (Bricker & Jones 1995). On the other hand, rivers play a major role in assimilation or transporting municipal and industrial wastewater and runoff from agricultural land. Municipal and industrial wastewater discharge constitutes a constant polluting source, whereas surface runoff is a seasonal phenomenon, largely affected by climate within the basin (Singh *et al.* 2004). Seasonal variations in precipitation, surface runoff, interflow, groundwater flow and pumped in and outflows have a strong effect on river discharge and, subsequently, on the concentration of pollutants in river water (Vega *et al.* 1998). Therefore, the effective long-term management of rivers requires a fundamental understanding of

doi: 10.2166/hydro.2008.008

hydro-morphological, chemical and biological characteristics. However, due to spatial and temporal variations in water quality (which are often difficult to interpret), a monitoring program, providing a representative and reliable estimation of the quality of surface waters, is necessary (Dixon & Chiswell 1996).

The application of different multivariate statistical techniques, such as 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, allows the identification of possible factors/sources that influence water systems and offers a valuable tool for reliable management of water resources as well as rapid solution to pollution problems (Helena *et al.* 2000; Lee *et al.* 2001; Adams *et al.* 2001; Wunderlin *et al.* 2001; Reghunath *et al.* 2002; Simeonova *et al.* 2003; Simeonov *et al.* 2004; Papatheodorou *et al.* 2007). Multivariate statistical techniques has been applied to characterize and evaluate surface and freshwater quality, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality. For example, Zou & Yu (1996) have used a general dynamic factor model to reduce the high dimensionality of the original matrix variables in order to detect trends in time series. Meng & Maynard (2001) used cluster and factor analysis to identify geochemical regions within a watershed in Brazil. Silva & Williams (2001) combined multivariate statistical analyses with GIS analysis to determine if a correlation existed between water quality and landscape characteristics in a watershed in southern Ontario (Canada). Singh *et al.* (2004, 2005) used multivariate statistical techniques to evaluate spatial and temporal variations in water quality of the Gomti River (India). Similarly, Shrestha & Kazama (2007) also used multivariate statistical techniques to assess the surface water quality of the Fuji River Basin in Japan. These techniques, especially PCA, have also been used to evaluate the water quality monitoring stations (Ouyang 2005). Furthermore, long-term hydrochemical data of shallow water bodies has been evaluated using factor analysis and discriminant analysis (Medina-Gomez & Herrera-Silverira 2003; Solidoro *et al.* 2004; Parinet *et al.* 2004).

In the present study, a large data matrix, obtained during a 6-year (1995–2000) monitoring program, is subjected to different multivariate statistical techniques to extract information about the similarities or 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 Mekong River.

STUDY AREA

The Mekong River (see Figure 1) is the longest in Southeast Asia and the twelvth longest in the world, with a length of 4800 km, a drainage area of 80,5604 km^2 (WRI *et al.* 2003) and an annual runoff of 475 billion m^3. It originates from the Tanggula Mountains at an altitude of about 5000 m a.s.l. on the Tibetan plateau, flows southwards through Tsinghai, Tibet and Yunnan provinces of China, Myanmar, Laos, Thailand, Cambodia and Vietnam, and discharges into the South China Sea. The upper Mekong River, called the Lancang River in China, flows by a series of giant mountain ranges, such as Ta Nian Ta Weng, Ning Jing, Heng Duan and Yun Ling. Therefore, the upper basin is extremely slender consisting of deep and narrow valleys, some as high as 1200 m above the river, e.g. the Lancang River in Tibet. In the upper 1600 km (from the Tibetan plateau to the Thailand–Myanmar border), the river drops rapidly by about 4500 m a.s.l., and then an additional 200 m a.s.l. along 830 km between the Thailand–Myanmar border and Vientiane in Laos.

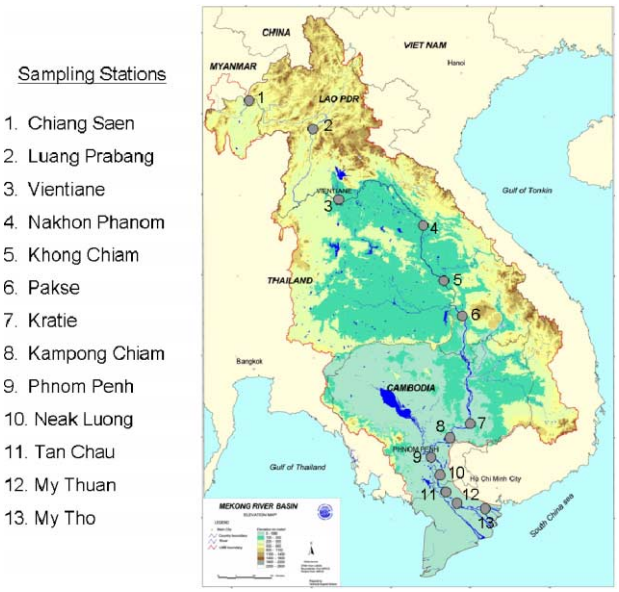

Figure 1 | Map of study area and water quality monitoring stations (listed 1–13) in the Mekong River.

After coming out of China and entering the Golden Triangle, it is called the Lower Mekong River and turns gentler, where most of it is navigable (Kite 2001). The climate of the upper Mekong basin is featured as the high mountainous cold weather in Tsinghai, Tibet and Yunnan in China. The lower Mekong basin is situated in the tropical zone and is dominated by two distinct monsoons: the rainy southwest and the dry northeast. The southwest monsoon from the Indian Ocean brings proper rainfall and lasts from mid-May to mid-October. During this period the basin experiences frequent rainfall (around 85–90% of annual rainfall), high humidity, maximum cloudiness and tropical temperatures. The northeast monsoon from China remains active from mid-October to April and causes a dry spell over the basin. The summer starts in mid-February and ends around mid-May. The weather in summer is rather hot with high temperature, low rainfall and low humidity. Lowest rainfall occurs in December and January. The flood season starts June–July and ends in November–December, with a peak in September, and accounts for 85–90% of the total annual runoff. During the dry season, the monthly flow accounts for only 1–2% of annual flow. Mean annual flow at different locations from upstream to downstream shows a great change, from Nong Khai (4490 m^3/s) to Kratie (13,700 m^3/s). Two general categories of soil have been identified in the MRB: (i) upland soils: podsols, red and black soils, lateritic and mountain soils and (ii) lowland soils: located in the Korat plateau in Northeast Thailand, the Mekong plain and delta. Theses soils are coastal complex, delta, flood plain and groundwater complex soils (MRC & UNEP 1997). Forestry, agriculture and fisheries are the major land uses of the Mekong River basin. Intrusion of saline water into the Mekong Delta and water quality degradation due to growing populations and increasing economic development are among the emerging issues regarding water usage in the Mekong River basin (MRC 2001).

MONITORED PARAMETERS AND ANALYTICAL METHODS

The datasets of 13 water quality monitoring stations which covers the Lower Mekong River of ~2400 km from its river mouth, comprising 18 water quality parameters monitored monthly as a grab samples over 6 years (1995–2000), were obtained from the Mekong River Commission (MRC). The selected water quality parameters includes pH, dissolved oxygen, chemical oxygen demand (manganese), calcium, magnesium, potassium, sodium, chloride, sulfate, iron, silicon, total suspended solids, electrical conductivity, alkalinity, nitrate nitrogen, ammonical nitrogen and inorganic dissolved phosphorus and total phosphorus. The water quality parameters, their units and the basic statistics of the monthly measured 6-year dataset on river water quality are summarized in Table 1.

METHODS

Data treatment 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 the suitability of the data for principal component analysis/factor analysis, Kaiser–Meyer–Olkin (KMO) and Bartlett's test were performed. KMO is a measure of sampling adequacy that indicates the proportion of variance which is common variance, i.e. which might be caused by underlying factors. High values (close to 1) generally indicate that principal component/factor analysis may be useful, which is the case in this study: KMO = 0.80. Bartlett's test of sphericity indicates whether the 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 significance relationships among variables.

The water quality parameters were grouped into four seasons: spring (March–May), summer (June–August), autumn (September–November) and winter (December–February), and each was assigned a numerical value in the data file (spring = 1; summer = 2; autumn = 3 and winter = 4), which, as a variable corresponding to the season, was correlated (pair by pair) with all the measured parameters.

River water quality datasets were subjected to three multivariate techniques: principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA). DA was applied to raw data, whereas PCA and FA 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

Table 1 | Means of different water quality parameters at different locations of the Mekong River during 1995–2000

	Region	Upstream			Midstream			Lower stream			Delta			
	Country	Thailand	Laos		Thailand		Laos	Cambodia			Cambodia	Vietnam		
S.N	Parameters	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6	St. 7	St. 8	St. 9	St. 10	St. 11	St. 12	St. 13
1	pH	7.88	7.76	7.93	7.76	7.77	7.78	7.38	7.58	7.54	7.48	7.42	7.35	7.23
2	TSS (mg l^{-1})	287.73	186.25	298.03	127.32	134.35	154.23	119.23	86.98	100.59	93.00	93.75	67.51	60.94
3	EC (μS m^{-1})	23.98	23.37	22.74	21.18	19.40	18.49	14.01	15.12	15.57	14.10	13.41	13.85	47.92
4	Ca (mg l^{-1})	1.41	1.54	1.48	1.21	1.18	1.15	0.77	0.85	0.87	0.77	0.73	0.75	0.86
5	Mg (mg l^{-1})	0.53	0.50	0.49	0.44	0.41	0.42	0.33	0.37	0.34	0.34	0.38	0.40	1.30
6	Na (mg l^{-1})	0.41	0.20	0.17	0.44	0.36	0.17	0.30	0.30	0.32	0.29	0.28	0.28	2.46
7	K (mg l^{-1})	0.05	0.03	0.03	0.05	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.12
8	Alk (mg l^{-1})	1.83	1.77	1.61	1.59	1.54	1.35	1.03	1.12	1.16	1.02	1.09	1.11	1.06
9	Cl (mg l^{-1})	0.25	0.16	0.19	0.30	0.23	0.25	0.14	0.15	0.16	0.15	0.21	0.18	2.78
10	SO_4 (mg l^{-1})	0.31	0.50	0.51	0.22	0.20	0.41	0.19	0.20	0.20	0.18	0.14	0.12	0.81
11	Fe (mg l^{-1})	0.05	0.03	0.04	0.04	0.07	0.02	1.14	0.34	0.15	0.37	0.62	0.97	1.43
12	NO_3-N (mg l^{-1})	0.33	0.17	0.16	0.32	0.28	0.10	0.14	0.16	0.18	0.18	0.25	0.31	0.41
13	NH_4-N (mg l^{-1})	0.05	0.03	0.04	0.04	0.04	0.03	0.03	0.04	0.03	0.04	0.01	0.02	0.05
14	PO_4-P (mg l^{-1})	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.06
15	TP (mg l^{-1})	0.06	0.03	0.05	0.05	0.05	0.03	0.02	0.03	0.03	0.03	0.12	0.10	0.24
16	Si (mg l^{-1})	5.09	6.33	6.52	4.95	5.00	5.98	5.33	4.85	4.83	4.37	5.84	5.97	5.99
17	DO (mg l^{-1})	8.07	3.91	6.81	7.51	8.06	3.90	7.32	6.70	6.74	6.24	6.69	6.65	6.53
18	COD (mg l^{-1})	1.79	1.24	1.11	1.34	1.38	1.25	2.03	1.94	1.83	2.28	2.56	2.20	2.93

the parameters analyzed. The z-scale transformation converts Pearson's r's to the normally distributed variable z. All mathematical and statistical computations were made using Microsoft Office Excel 2003 and STATISTICA 6.

Principal component analysis/factor analysis

PCA is designed to transform the original variables into new, uncorrelated variables (axes), called the principal components, which are linear combinations of the original variables. The new axes lie along the directions of maximum variance. 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 & Pop 2005). PC provides information on the most meaningful parameters, which describes a whole dataset affording data reduction with minimum loss of original information (Helena *et al.* 2000). The principal component (PC) can

be expressed as

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + a_{i3}x_{3j} + \ldots + a_{im}x_{mj} \quad (1)$$

where z is the component score, a is the component loading, x is the measured value of variable, i is the component number, j is the sample number and m is the total number of variables.

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. This purpose can be achieved 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 (Vega *et al.* 1998; 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. As a result, a small number of factors will usually account for approximately the same amount of information as does the much larger set of original observations. The FA can be expressed as

$$z_{ji} = a_{f1}f_{1i} + a_{f2}f_{2i} + a_{f3}f_{3i} + \ldots + a_{fm}f_{mi} + e_{fi} \quad (2)$$

where z is is the component score, a is the the component loading, f is the factor score, e is the residual term accounting for errors or other source of variation, i is the sample number and m is the total number of variables.

Discriminant analysis

Discriminant analysis (DA) is used to classify cases into categorical-dependent values, usually a dichotomy. If discriminant analysis 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. 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. The DA technique builds up a discriminant function for each group, which operates on raw data (Johnson & Wichern 1992; Wunderlin *et al.* 2001; Singh *et al.* 2004, 2005), as in the equation below:

$$f(G)i = k_i + \sum_{j=1}^{n} w_{ij}p_{ij} \quad (3)$$

where i is the number of groups (G), k_i is the constant inherent to each group, n is the number of parameters used to classify a set of data into a given group, w_j is the weight coefficient, assigned by DA to a given selected parameter (P_j). The weight coefficient maximizes the distance between the means of the criterion (dependent) variable. The classification table, also called a confusion, assignment or prediction matrix or table, is used to assess the performance of DA. This is simply a table in which the rows are the observed categories of the dependents and the columns are the predicted categories of the dependents. When prediction is perfect, all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

In this study, DA was performed on each raw data matrix using standard, forward stepwise and backward stepwise modes in constructing discriminant functions 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 DISCUSSIONS

Temporal variations in water quality

Temporal DA was performed on raw data after dividing the whole dataset into four seasonal groups (spring, summer, autumn and winter). The temporal variations in water quality were evaluated between monitoring regions and within monitoring regions (upstream (UR), midstream (MR), lower stream (LR) and delta (DR) regions). The upstream and the midstream consists of the water quality monitoring stations in Thailand and Laos; the lower stream consists of stations in Cambodia and delta consists of stations in Cambodia and Vietnam. Discriminant functions (DFs) and classification matrices (CMs) were obtained from the standard, forward

stepwise and backward stepwise modes of DA. 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. The standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 and 17 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 and 17 discriminant variables, respectively, yielded the corresponding CMs, assigning 53% of the cases correctly. However, in backward stepwise mode, DA gave CMs with 46% correct assignations using only three discriminant parameters with little difference in the match for each season compared with the forward stepwise mode. Thus, the temporal DA results suggest that electrical conductivity, alkalinity and nitrate nitrogen are the most significant parameters to discriminate between the four seasons, which means that these three parameters account for most of the expected temporal variations in the river water quality between monitoring regions of the Mekong River basin.

Temporal variations in river water quality were also performed within the monitoring regions. In the upstream region of the Mekong River (UR), the standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 and 10 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 and 10 discriminant variables, respectively, yielded the corresponding CMs, assigning 68% of the cases correctly. However, in backward stepwise mode, DA gave CMs with 63% correct assignations using only two discriminant parameters with little difference in the match for each season compared with the forward stepwise mode. Thus, the temporal DA results suggest that electrical conductivity and silicon are the most significant parameters to discriminate between the four seasons, which means that these two parameters account for most of the expected temporal variations in the river water quality in the upstream region of the Mekong River basin. In the midstream region of the Mekong River (MR), the standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 and 8 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 and 8 discriminant variables, respectively, yielded the corresponding CMs, assigning 62% of the cases correctly. However, in backward stepwise mode, DA gave CMs with 55% correct assignations using only three discriminant parameters with little difference in the match for each season compared with the forward stepwise mode. Thus, the temporal DA results suggest that total suspended solids, electrical conductivity and silicon are the most significant parameters to discriminate between the four seasons, which means that these three parameters account for most of the expected temporal variations in the river water quality in the midstream of Mekong River basin. In the lower stream region of the Mekong River (LR), the standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 and 12 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 and 12 discriminant variables, respectively, yielded the corresponding CMs, assigning 79% of the cases correctly. However, in backward stepwise mode, DA gave CMs with 65% correct assignations using only three discriminant parameters with little difference in the match for each season compared with the forward stepwise mode. Thus, the temporal DA results suggest that electrical conductivity and sodium are the most significant parameters to discriminate between the four seasons, which means that these two parameters account for most of the expected temporal variations in the river water quality in the lower stream region of the Mekong River basin. In the delta region (DR), the standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 and 15 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 and 15 discriminant variables, respectively, yielded the corresponding CMs, assigning 63% of the cases correctly. However, in backward stepwise mode, DA gave CMs with 65% correct assignations using only two discriminant parameters with little difference in the match for each season compared with the forward stepwise mode. Thus, the temporal DA results suggest that total suspended solids and electrical conductivity are the most significant parameters to discriminate between the four seasons, which means that these two parameters account for most of the expected temporal variations in the river water quality in the delta region of the Mekong River basin.

As identified by DA, box and whisker plots of the selected parameters showing seasonal trends are given in Figure 2. The temporal DA results suggest that electrical conductivity, alkalinity and nitrate nitrogen account for most of the expected temporal variations in the river water quality between monitoring regions of the Mekong River basin. The electric conductivity and alkalinity are lower in the summer

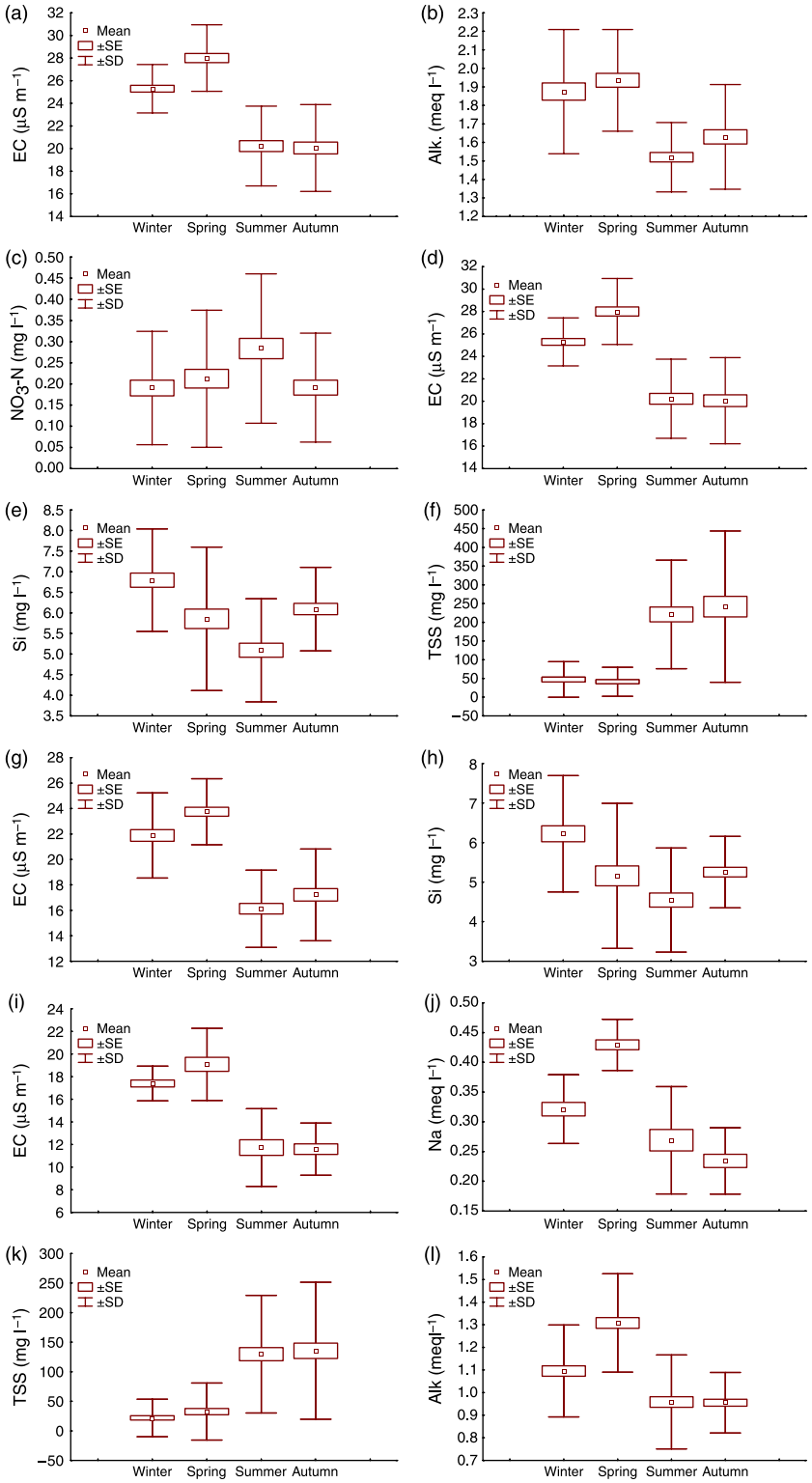

Figure 2 | Temporal variations in water quality of the Mekong River: (a)–(c) between monitoring regions; (d), (e) in upstream region; (f)–(h) midstream region; (i), (j) lower stream region; (k), (l) delta region. (EC = electrical conductivity; iron, Si = silicon, TSS = total suspended solids, Alk = alkalinity, NO_3-N = nitrate nitrogen, Na = sodium).

and autumn seasons, which reflects the dilution effect. However, the concentration of nitrate nitrogen is higher in the summer season, which can be attributed to the non-point source pollution, i.e. agricultural runoff. In the upper region, electrical conductivity and silicon are lower in the summer and autumn seasons as compared to the winter and spring seasons, which also suggest the dilution effect. In the midstream region also electric conductivity and silicon are lower but total suspended solids are higher in these seasons, which suggest erosion and transport from riparian areas. In the lower stream region, electrical conductivity and sodium are lower in the summer and autumn seasons, which suggest the dilution effect. In the delta region, total suspended solids and alkalinity are lower in the summer and autumn seasons as compared to the winter and spring seasons.

Spatial variations in water quality

Spatial DA was performed with the same raw dataset comprising 18 parameters after grouping into four major classes of monitoring regions (UR, MR, LR and DR) and countries (Thailand, Laos, Cambodia and Vietnam). The monitoring regions and countries were the grouping (dependent) variable, while all the measured parameters constituted the independent variables. Discriminant functions (DFs) and classification matrices (CMs) were obtained from the standard, forward stepwise and backward stepwise modes of DA. Similarly to temporal DA, the standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 parameters. Both the standard and forward stepwise mode DFs using 18 discriminant parameters yielded the corresponding CMs, assigning more than 71% cases correctly, whereas the backward stepwise mode DA gave CMs with 69% correct assignations using only eight discriminant parameters. Backward stepwise DA shows that total suspended solids, calcium, sodium, alkalinity, chloride, iron, nitrate nitrogen and total phosphorus are the discriminating parameters in between the monitoring regions.

Box and whisker plots of discriminating parameters identified by spatial DA (backward stepwise mode) were constructed to evaluate different patterns associated with variations in river water quality between the monitoring regions (Figure 3). The decrease in total suspended solids, alkalinity and calcium is observed from the upstream region to the delta region. The higher concentration of sodium and chloride is observed in the delta region. Similarly a higher concentration of iron is observed in the lower stream and delta regions. The higher concentration of nitrate nitrogen and total phosphorus is observed in the delta region.

The standard DA mode- and forward stepwise DA mode-constructed DFs, including 18 parameters, respectively. Both the standard and forward stepwise mode DFs using 18 discriminant parameters, respectively, yielded the corresponding CMs assigning 80% of cases correctly, whereas the backward stepwise mode DA gave CMs with 79% correct assignations using only 12 discriminant parameters. Backward stepwise DA shows that total suspended solids, calcium, sodium, potassium, alkalinity, chloride, sulfate, iron, nitrate nitrogen, total phosphorus, silicon and dissolved oxygen are the discriminating parameters between the countries.

Box and whisker plots of discriminating parameters identified by spatial DA (backward stepwise mode) were constructed to evaluate different patterns associated with variations in river water quality between the countries (Figure 4). The concentrations of total suspended solids, calcium, silicon and sulfate are comparatively higher in Laos as it includes one sampling station, Chiang Saen, upstream of Thailand. The concentrations of alkalinity and dissolved oxygen are comparatively higher in Thailand. The concentrations of sodium, chloride, potassium, iron, nitrate nitrogen and total phosphorus are comparatively higher in Vietnam. The results of the spatial variations can be used to select the polluted areas and set the priority areas for the river water quality management in the study area.

Data structure determination and source identification

Principal component analysis/factor analysis was performed on the normalized datasets (18 variables) separately for individual sampling stations and for the whole Mekong River to compare the compositional pattern between analyzed water samples and identify the factors influencing each one. The input data matrices (variables × cases) for PCA/FA were [18 × 72] for each sampling stations and [18 × 855] for the whole river. PCA of the 13 datasets of the individual stations yielded four PCs for station 9 (Phnom Penh); five PCs for station 1 (Chiang Saen), station 2 (Luang Prabang), station 5 (Khong Chiam), station 8 (Kampong Chiam) and station 10

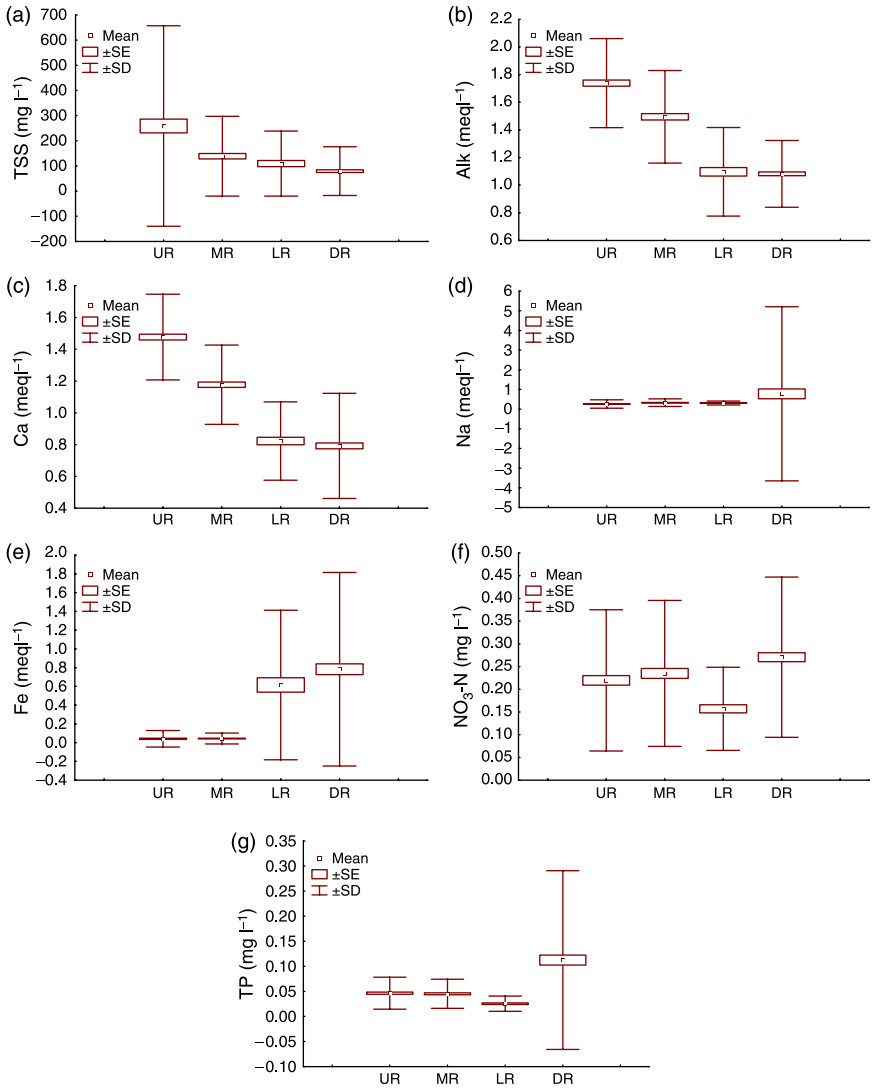

Figure 3 | Spatial variations: (a) TSS, (b) Alk., (c) Ca, (d) Na, (e) Fe, (f) NO_3 – N and (g) TP river water quality between monitoring regions of the Mekong River.

(Neak Luong); six PCs for station 3 (Vientine), station 4 (Nakhon Phanom), station 6 (Pakse), station 7 (Kratie) and station 12 (Mythuan); seven PCs for station 13 (Mytho); and eight PCs for station 11 (Tan Chau) with eigenvalues > 1, explaining 68, 66, 73, 68, 71, 69, 68, 73, 72, 77, 65, 80 and 71% of the total variance in the respective water quality datasets. Similarly, PCA of the whole Mekong River dataset yielded six PCs with eigenvalues > 1, explaining 72% of the variance in the water quality datasets. 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 & Muller 1987). Equal numbers of vari-factors (VFs) were obtained for all cases through FA performed on the PCs. Corresponding VFs, variable loadings and explained variance are presented in Table 2. 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.

In the case of the whole river, the VF1, which explained 26.3% of the whole dataset, has strong positive loadings on EC, sodium and chloride. This factor can be named as the salinity factor. There are two reasons for the salinity problem in the Mekong River. The first, mainly in Northeast Thailand (upstream and midstream), has resulted from deforestation

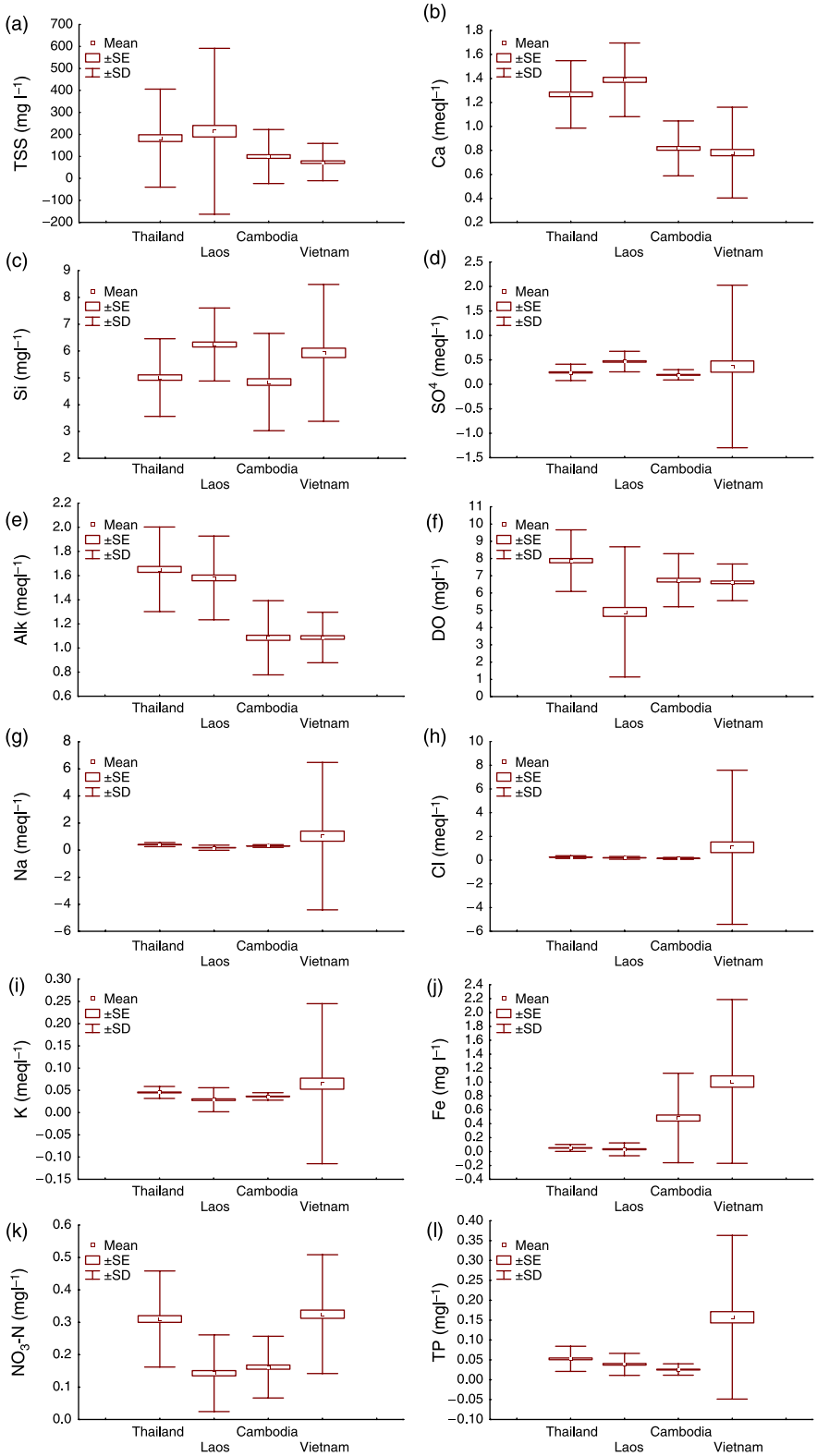

Figure 4 | Spatial variations: (a) TSS, (b) Ca, (c) Si, (d) Alk., (e) DO, (f) Na, (g) Cl, (h) K, (i) Fe, (j) Fe, (k) NO_3 – N and (l) TP river water quality between countries of the Mekong River.

Table 2 | Water quality variables associated with strong factor loadings, variance explained by each factors and total variance explained by the datasets of the whole Mekong River and individual sampling stations

Location	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Total variance (%)
Whole River	EC, Na, Cl	PH, Ca, Alk	Mg, K, SO_4	TP, PO_4	TSS	DO	–	–	72
E.V	4.73	3.07	1.44	1.42	1.11	1.10	–	–	
Variance (%)	26.30	17.05	8.00	7.89	6.18	6.12	–	–	
Chang Saen	EC, Ca, Na, Alk, K, Cl	TP	pH	Si	Fe	–	–	–	66
E.V	5.89	1.76	1.53	1.49	1.24	–	–	–	
Variance (%)	32.70	9.79	8.49	8.25	6.87	–	–	–	
Luang Prabang	EC*, Ca*, SO_4^*	Na, K	TP, PO_4	pH	NH_4^*	–	–	–	73
E.V	4.36	2.63	2.20	1.51	1.24	–	–	–	
Variance (%)	24.24	14.62	12.22	8.39	6.91	–	–	–	
Vientiane	Ca, Mg	SO_4^*	Na, K	DO	TP*, PO_4^*	EC, Alk	–	–	68
E.V	4.15	2.35	1.98	1.44	1.23	1.03	–	–	
Variance (%)	23.06	13.07	11.00	8.00	6.81	5.70	–	–	
Nakhon Phanom	EC, Ca, Mg, Alk	Na, K, Cl	NO_3	Si*, DO*	PO_4^*	TP*	–	–	73
E.V	5.71	2.39	1.62	1.30	1.05	1.01	–	–	
Variance (%)	31.70	13.28	9.03	7.25	5.86	5.59	–	–	
Khong Chiam	PH, TSS*, EC, Ca, Mg, Alk, COD_{Mn}^*	Si	Fe	DO*	Cl	–	–	–	68
E.V	6.23	1.71	1.65	1.53	1.08	–	–	–	
Variance (%)	34.63	9.49	9.18	8.48	6.02	–	–	–	
Pakse	TSS*, EC, Ca, Alk	Na, K	TP, PO_4	NO_3^*, COD_{Mn}^*	Si	Fe	–	–	72
E.V	4.85	2.40	1.91	1.39	1.31	1.04	–	–	
Variance (%)	26.93	13.31	10.64	7.70	7.26	5.77	–	–	
Kratie	EC, Ca, Alk, Cl, Fe*	pH*, COD_{Mn}^*	K	Si	TP*, PO_4^*	NO_3^*	–	–	77
E.V	7.07	1.72	1.53	1.24	1.21	1.05	–	–	
Variance (%)	39.28	9.55	8.49	6.91	6.74	5.84	–	–	
Kampong Chiam	EC, Ca, Mg, SO_4	COD_{Mn}	K	Si	NO_3^*	–	–	–	71
E.V	6.86	1.98	1.50	1.36	1.05	–	–	–	
Variance (%)	38.09	11.00	8.36	7.55	5.81	–	–	–	
Phnom Penh	TSS, EC*, Ca*,Mg*, Alk*,Cl*, TP, PO_4	K	Fe*	DO*	–	–	–	–	71
E.V	7.31	2.23	1.64	1.53	–	–	–	–	
Variance (%)	40.60	12.39	9.12	8.48	–	–	–	–	
Neak Leang	EC, Ca, Na, Mg, Alk, Cl, SO_4	NO_3	Si	Fe*	NH_4	–	–	–	69

Table 2 | (*continued*)

Location	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Total variance (%)
E.V	6.42	1.95	1.51	1.35	1.11	–	–	–	
Variance (%)	35.69	10.83	8.38	7.51	6.18	–	–	–	
Tan Chau	PH, Ca	Na, K, Cl	NH_4	DO*	Mg	PO_4^*	COD_{Mn}^*	Fe*	71
E.V	3.32	2.41	1.87	1.24	1.15	1.09	1.05	1.01	
Variance (%)	18.44	13.40	10.38	6.87	6.39	6.06	5.83	5.61	
Mythuan	EC, Na	CA, Mg, Alk	NO_3	TSS, PO_4	NH_4, Si*	SO_4	–	–	65
E.V	3.54	2.36	1.76	1.53	1.31	1.14	–	–	
Variance (%)	19.66	13.12	9.75	8.50	7.27	6.32	–	–	
Mytho	EC, Ca, Na, Cl	Mg, K, SO_4	TP, PO_4	DO*	TSS	NO_3^*	Si*	–	80
E.V	5.78	1.86	1.67	1.50	1.32	1.23	1.06	–	
Variance (%)	32.11	10.34	9.30	8.35	7.34	6.83	5.87	–	

E.V = eigenvalue, * negative loadings

and associated rises in water tables. The problem has been exacerbated where there are underlying salt domes. The second set of salinity problems arises in the delta, where saline intrusion occurs during the dry season as insufficient fresh water flows out to keep tidal seawater at bay (Hirsch & Cheong 1996).

In the upstream region, the VF1 of station 1 and station 2, which has strong positive loadings of mineral salts, explained the 32.70% and 24.24% variance of the whole dataset. But the VF1 of Vientine has strong positive loadings of hardness component and explained 23.06% of variance. This factor accounts for the temporary hardness of water.

In the middle stream region, all stations (station 4, station 5 and station 6) have strong positive loadings of mineral salts and hardness components. Moreover, in station 5 VF1 is also associated with organic matter loadings. The negative correlation between COD_{Mn} and pH is due to anaerobic conditions in the river from the loading of high dissolved organic matter, which results in formation of ammonia and organic acids, leading to a decrease in pH.

In the lower stream region, VF1 is associated with strong loadings of mineral salts composition, weathering and anthropogenic source of pollution. In station 7, apart from mineral salts, iron also has strong loadings. This station receives water from three other tributaries, Se Kong, Se Sang and Sre Pok. The highest loadings of iron can be attributed to the soils containing pyrite (iron sulfide). The loading of iron is also observed in stations located in the lower stream and delta regions either as a single factor or associated with other variables. Pyrite is usually found in deltaic areas where rising sea levels have flooded mangroves and estuaries (Hirsch & Cheong 1996). In station 9, VF1 has strong loadings of TP and PO_4 which indicates the influence of an anthropogenic source of pollution. Phnom Penh is a transition area where the Mekong River drains water into the Great Lake through the Tonle Sap River during the flood season, and as the water level in the Mekong goes down water drains out from the Lake into the Mekong River and the Bassac River, one of the two main branches of the Mekong River. The reversal of the flow of the Tonle Sap River is a unique hydrological phenomena that ensures a higher dry season flow for the delta than if it only received water from the Mekong River. This causes the year-round mixing of sewage water from Bassac River and Tonle Sap River areas (Sokha 2004).

In the delta region (station 10, station 12 and station 13), the VF1 is associated with strong loadings of mineral salts. This is due to saline intrusion occuring during the dry season as insufficient fresh water flows out to keep tidal seawater at bay (Hirsch & Cheong 1996).

CONCLUSIONS

In this study, multivariate statistical techniques such as discriminant analysis (DA), principal component analysis (PCA) and factor analysis (FA) were used to evaluate spatial and temporal variations in surface water quality of the Mekong River using six years (1995–2000) datasets of 18 water quality variables covering 13 sampling stations. The analysis was conducted for a whole river including all stations and dividing the whole Mekong River into four regions: upstream region (UR), midstream region (MR), lower stream region (LR) and delta region (DR). Discriminant analysis was used to find out the spatial and temporal variation in river water quality. Approximately 50–80% correctness was obtained in the classification matrix while performing discriminant analysis in this study. Spatial discriminant analysis revealed 8 parameters (total suspended solids, alkalinity, calcium, sodium, chloride, iron, nitrate nitrogen, total phosphorus) and 12 parameters (total suspended solids, alkalinity, calcium, sodium, potassium, chloride, sulfate, iron, nitrate nitrogen, total phosphorus, silicon, dissolved oxygen) are responsible for significant variations between monitoring regions and countries, respectively. Temporal discriminant analysis revealed 3 parameters (electrical conductivity, alkalinity, nitrate nitrogen) between monitoring regions; 3 parameters (total suspended solids, electrical conductivity, silicon) in the middle region; and 2 parameters (electrical conductivity, silicon) in the upper, lower and delta regions are the most significant parameters to discriminate between the four different seasons (spring, summer, autumn, winter). Therefore, DA allowed a reduction in the dimensionality of the large dataset, delineating a few indicator parameters responsible for large variations in water quality. Although the PCA/FA did not result in a significant data reduction as compared to DA, it helped to extract and identify the factors/sources responsible for variations in river water quality at different sampling sites. The results of PCA/FA, while considering the only factors that explain the highest variance in the dataset and with the highest eigenvalues revealed that most of the variations are explained by dissolved mineral salts along the whole Mekong River and in individual stations as well. But, in two sites (Khong Chiam and Phnom Penh) organic components (chemical oxygen demand) and nutrients (phosphorus) are also included. Thus, this study illustrates the usefulness of multivariate statistical techniques for analysis and interpretation of complex datasets, and in water quality assessment, identification of pollution sources/factors and understanding temporal/spatial variations in water quality for effective river water quality management.

ACKNOWLEDGEMENTS

The authors sincerely thank Pekh Sokhem for requesting data and staff of the Mekong River Commission (MRC) for providing the database and the Fuji Xerox Setsutaro Kobayashi Memorial Fund for providing funding support. We would also like to acknowledge the help and support provided by the 21st Century Center of Excellence (COE), Integrated River Basin Management in Asian Monsoon Region, University of Yamanashi.

REFERENCES

Adams, S., Titus, R., Pietesen, K., Tredoux, G. & Harris, C. 2001 Hydrochemical characteristic of aquifers near Sutherland in the Western Karoo. *South Africa. J. Hydrol.* **241**, 91–103.

Bricker, O. P. & Jones, B. F. 1995 Main factors affecting the composition of natural waters. In *Trace Elements in Natural Waters* (ed. B. Salbu & E. Steinnes), CRC Press, Boca Raton, FL, pp. 1–5.

Dixon, W. & Chiswell, B. 1996 Review of aquatic monitoring program design. *Wat. Res.* **30**, 1935–1948.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernández, J. M. & Fernández, L. 2000 Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Wat. Res.* **34**, 807–816.

Hirsch, P. & Cheong, G. 1996 *Natural Resource Management in the Mekong Basin: Perspectives for Australian Development Cooperation*. Final overview report to AusAID, University of Sydney.

Johnson, R. A. & Wichern, D. W. 1992 *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

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*. Sage University Press, Newbury Park.

Kite, G. 2001 Modeling the Mekong: hydrological simulation for environmental impact studies. *J. Hydrol.* **253**, 1–13.

Lee, J. Y., Cheon, J. Y., Lee, K. K., Lee, S. Y. & Lee, M. H. 2001 Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons. *J. Environ. Qual.* **30**, 1548–1563.

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. *Sci. Total Environ.* **313**, 77–89.

Medina-Gomez, I. & Herrera-Silverira, J. A. 2003 Spatial characterization of water quality in a karstic coastal lagoon without anthropogenic disturbance: a multivarate approach. *Estuarine Coastal Shel Sci.* **58**, 455–465.

Meng, S. X. & Maynard, J. B. 2001 The use of statistical analysis to formulate conceptual models of geochemical behavior: water chemical data from the Botucatu Aquifer in the Sao Paulo State, Brazil. *J. Hydrol.* **250**, 78–97.

MRC & UNEP 1997 *Mekong River Basin Diagnostic Study*. Final Report. Mekong River Commission (MRC), Bangkok, Thailand and United Nations Environment Programme (UNEP).

MRC 2001 *MRC Hydropower Development Strategy. Meeting the Needs, Keeping the Balance*. Mekong River Commission (MRC), Phnom Penh, Cambodia.

Ouyang, Y. 2005 Evaluation of river water quality monitoring stations by principal component analysis. *Wat. Res.* **39**, 2621–2635.

Papatheodorou, G., Lambrakis, N. & Panagopoulos, G. 2007 Application of multivariate statistical procedures to the hydrochemical study of coastal aquifer: an example from Crete, Greece. *Hydrol. Process.* **21** (11), 1482–1495.

Parinet, B., Lhote, A. & Legube, B. 2004 Pricipal component analysis: an appropriate tool for water quality evaluation and management - application to a tropical lake system. *Ecol. Modell.* **178**, 295–311.

Reghunath, R., Murthy, T. R. S. & Raghavan, B. R. 2002 The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. *Wat. Res.* **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.

Shrestha, S. & Kazama, F. 2007 Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji River basin, Japan. *Environ. Modell. Soft.* **22** (4), 464–475.

Silva, L. & Williams, D. D. 2001 Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Wat. Res. Res.* **35** (14), 3372–3462.

Simeonov, V., Simeonova, P. & Tsitouridou, R. 2004 Chemometric quality assessment of surface waters: two case studies. *Chem. Engng. Ecol.* **11** (6), 449–469.

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. & Kouimtzis, T. 2003 Assessment of the surface water quality in Northern Greece. *Wat. Res.* **37**, 4119–4124.

Simeonova, P., Simeonov, V. & Andreev, G. 2003 Environmetric analysis of the Struma River water quality. *Central Europ. J. Chem.* **2**, 121–126.

Singh, K. P., Malik, A., Mohan, D. & Sinha, S. 2004 Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. *Wat. Res.* **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. *Anal. Chim. Acta* **538**, 355–374.

Sokha, C. 2004 Ecological risk assessment training project. In: *Phnom Penh System. Regional Workshop on Water Quality and Environmental Monitoring, Vientiane, Lao PDR, 15–16 November*.

Solidoro, C., Pastres, R., Cossarini, G. & Ciavatta, S. 2004 Seasoal and spatial variability of water quality parameters in the lagoon of Venice. *J. Marine Syst.* **51**, 7–18.

Vega, M., Pardo, R., Barrado, E. & Deban, L. 1998 Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Wat. Res.* **32**, 3581–3592.

WRI, IUCN, IWMI & the Ramsar Convention Bureau 2003 *The Watersheds of the World CD*. World Resources Institute, Washington, DC. Available at: http://www.waterandnature.org/eatlas.

Wunderlin, D. A., Diaz, M. P., Ame, M. V., Pesce, S. F., Hued, A. C. & Bistoni, M. A. 2001 Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). *Wat. Res.* **35**, 2881–2894.

Zou, S. & Yu, Y. S. 1996 A dynamic factor model for multivariate water quality of time series with trends. *J. Hydrol.* **178**, 381–400.

First received 29 June 2006; accepted in revised form 4 July 2007