

Application of the chemometric approach to evaluate the spatial variation of water chemistry and the identification of the sources of pollution in Langat River, Malaysia

Wan Ying Lim · Ahmad Zaharin Aris ·
Sarva Mangala Praveena

Received: 24 February 2012 / Accepted: 31 October 2012 / Published online: 10 November 2012
© Saudi Society for Geosciences 2012

Abstract The current study presents the application of selected chemometric techniques—hierarchical cluster analysis (HCA) and principal component analysis (PCA)—to evaluate the spatial variation of the water chemistry and to classify the pollution sources in the Langat River. The HCA rendered the sampling stations into two clusters (group 1 and group 2) and identified the vulnerable stations that are under threat. Group 1 (LY 1 to LY 14) is associated with seawater intrusion, while group 2 (LY 15 to LY 30) is associated with agricultural and industrial pollution. PCA analysis was applied to the water datasets for group 1 resulting in four components, which explained 85 % of the total variance while group 2 extracted six components, explaining 88 % of the variance. The components obtained from PCA indicated that seawater intrusion, agricultural and industrial pollution, and geological weathering were potential sources of pollution to the study area. This study demonstrated the usefulness of the chemometric techniques on the interpretation of large complex datasets for the effective management of water resources.

Keywords Hierarchical cluster analysis · Principal component analysis · Surface water

Introduction

In recent years, increasing attention has been given to surface water quality. The quality of surface water is an essential component of the natural environment and is considered as the main factor for controlling environmental health and potential hazards. Studies have shown that the quality of surface water is commonly determined by both natural and anthropogenic influences, including catchment geology, atmospheric inputs, anthropogenic inputs, and climatic conditions (Shrestha and Kazama 2007; Altin et al. 2009; Saim et al. 2009; Shokrzadeh and Saedi Saravi 2009; Najjar and Khan 2012). Due to intensive human activities, the anthropogenic inputs from urban, mining, industrial, and agricultural activities are the primary factors affecting the surface water quality. As such, surface water including rivers, lakes, estuaries, and seas are most susceptible to pollution owing to their easy accessibility for wastewater disposal especially for those close to highly urbanized regions (Singh et al. 2004). Therefore, a monitoring program that is capable of presenting a reliable estimation of the quality of surface water is necessary in order to evaluate the spatial and temporal variations. Comparing the value of the environmental variables with existing guidelines is the most common method in water quality assessment. However, this method does not readily give information regarding the status of the pollution sources (Debels et al. 2005).

Environmental datasets are usually complex and contain a large amount of information with internal relationships among variables, often in a partially hidden structure (Saim et al. 2009; Praveena et al. 2011). Indeed, the large and complicated data matrix results from surface water monitoring programs will lead to difficulty in interpretation and evaluation of the observed

W. Y. Lim · A. Z. Aris (✉)
Environmental Forensics Research Centre,
Faculty of Environmental Studies,
Universiti Putra Malaysia, 43400, UPM,
Serdang, Selangor, Malaysia
e-mail: zaharin@env.upm.edu.my

S. M. Praveena
Department of Environmental and Occupational Health,
Faculty of Medicine and Health Sciences,
Universiti Putra Malaysia, 43400, UPM,
Serdang, Selangor, Malaysia

quality data. Thus, the analysis of such complex data requires statistical techniques, especially chemometric studies (multivariate statistical techniques) to assess the water quality with respect to sustainability (Simeonov et al. 2004; Singh et al. 2004). Multivariate statistical techniques are powerful tools for analyzing a large number of datasets, classifying similarity and assessing the impact of humans on the water quality and ecosystem conditions (Shrestha and Kazama 2007; Praveena et al. 2011). This technique has been widely applied to a variety of environmental applications, which suggests its versatility in handling various types of data (Shrestha and Kazama 2007; Praveena et al. 2008; Krishna et al. 2009). There are comparable studies including the evaluation of temporal/spatial variations and the interpretation of water quality on the surface water (Shrestha and Kazama 2007; Hussain et al. 2008; Kazi et al. 2009; Krishna et al. 2009; Najjar and Khan 2012; Varol et al. 2012) and groundwater (Aris et al. 2007; Krishna et al. 2009; Aris et al. 2012). Hence, it can be considered that the multivariate technique provides a valuable way for the reliable management of water resources through the identification of the possible sources that influence the water system (Kazi et al. 2009).

Hierarchical cluster analysis (HCA) and principal component analysis (PCA) have been frequently applied to analyze the similarities among the sampling sites and source apportionment of pollution parameters in surface water (Akbal et al. 2011; Aris et al. 2012). HCA coupled with PCA is a powerful pattern recognition technique. HCA can be used to explain the interrelations among variables or sampling sites (Reghunath et al. 2002; Singh et al. 2005) and group the objects of interest into clusters based on the similarity within a class and dissimilarities among different classes (Bu et al. 2010; Praveena et al. 2011). In comparison, PCA is frequently employed in hydrochemical studies, geology and hydrogeology applications (Akbal et al. 2011; Najjar and Khan 2012). It is a dimension-reduction technique that provides information on the most significant parameters with a simpler representation of the data, as well as a reduction in the memory required and faster classification (Shrestha and Kazama 2007). The potential factors or sources that affect water systems can be identified by reducing the dimensionality of the dataset (Davis 1986; Huang et al. 2011; Praveena et al. 2011). The objective of this study is to investigate the spatial variations in surface water quality and to identify the potential sources of pollution of the Langat River. Chemometric methods (HCA and PCA) were used to evaluate the information concerning the similarities between the sampling stations and to ascertain the contribution of the potential factors or pollution sources among 29 parameters at 30 different sampling points of the Langat River. Based on the information obtained, a holistic interpretation of the results and the use of selected parameters as source tracers for contamination were enhanced.

Materials and methods

Site descriptions

The Langat River Basin occupies the south and southeastern parts of the state of Selangor in Malaysia (Fig. 1). The basin lies between latitudes $2^{\circ} 40' 152''\text{N}$ to $3^{\circ} 16' 15''\text{N}$ and longitudes $101^{\circ} 19' 20''\text{E}$ to $102^{\circ} 1' 10''\text{E}$ with a total catchment area of approximately $1,815 \text{ km}^2$. The basin can be divided into three areas: the mountainous area, the hilly area, and the lowland area (DOA 1995). The main river course is 141 km long. The river flows from the high hills in the north towards the plains and turns westward towards the coast of the state of Selangor (Mokhtar et al. 2009). The major tributaries are the Semenyih River, the Labu River, and the Mantin River. The Langat River is essential to the Selangor population and serves as one of the most important freshwater ecosystems in Selangor. Besides providing potable water, Langat River also supplies water for manufacturing and agricultural production. There are two major impoundments (Langat Dam and Semenyih Dam) that supply water to the entire basin. The source of the Langat River is on the Pahang–Selangor border where the hilly terrain reaches up to 1,500 m above the mean sea level. The basin consists of two estuaries, one is located on the northeastern side and the river water flows into the Lumut Strait while the other is on the southern side and flows directly into the Strait of Malacca (Mokhtar et al. 2009). Water samples were collected from different sampling sites covering from Dengkil to these two estuaries.

The Langat River receives an annual rainfall of 1,500 to 2,900 mm. The basin experiences an average temperature of 32°C throughout the year with a relative humidity of approximately 80 %. The basin is underlain by schist, phyllite, and granite rock formation of Permian age. The bedrock in the mountainous area includes Permian igneous rocks, Pre-Devonian schist, and phyllite of the Howthornden Formation (Gobbett and Hutchison 1973). The bedrock in the hilly area is predominantly Permo-Carboniferous meta-sandstone, consisting of mainly quartzite and slates of Kajang Formation and Kenny Hill Formation (Gobbett and Hutchison 1973; Taha 2003). The lowland area is of Quaternary deposits of Beruas, Gula, and Simpang Formations. These formations overlie the sedimentary bedrock of the Kenny Hill Formation and Kajang Hill Formation and grow progressively younger and thicker toward the coast (Gobbett and Hutchison 1973). The Quaternary deposits are made up of marine and continental deposits, which consist of gravel, sand, clay, and silt (JICA and MGDM 2002; Taha 2003).

Field sampling and preservation

The sampling was carried out in the rainy season (December 2010). The intense rainfall during the rainy season erodes

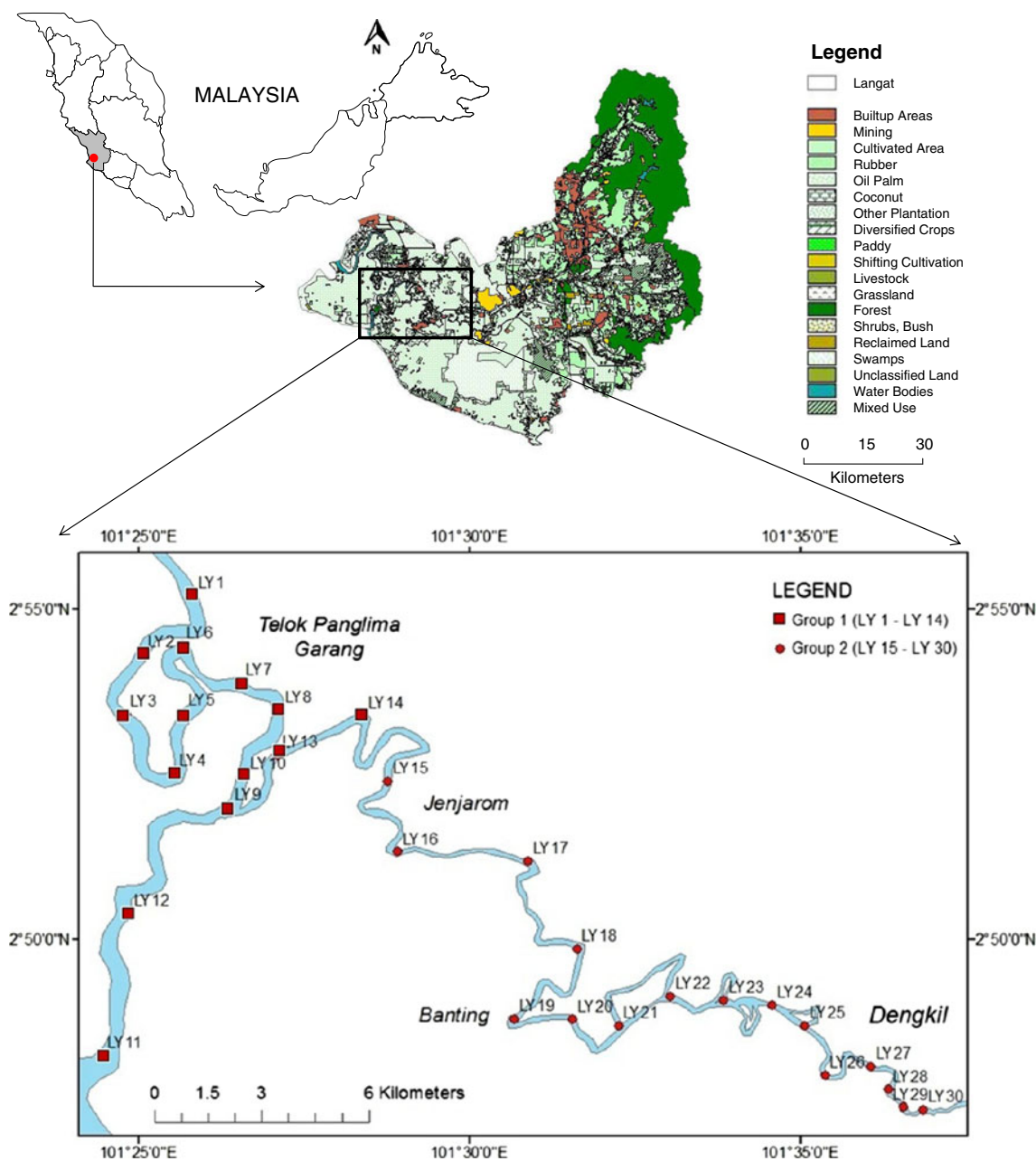


Fig. 1 Map of sampling stations in Langat River

the topsoil and carries the accumulated pollutants by surface runoff before draining into the river. Consequently, the elevated concentrations of certain pollutants are more likely to be evident in the rainy season. Triplicate water samples were collected from 30 sampling stations and were homogenized. During sampling, the pre-cleaned polyethylene bottles that were used were normalized by rinsing thoroughly with the river water to be collected and filled with running water facing the direction of the flow. In order to prevent the occurrence of biochemical and surface reaction of the water samples during transportation and storage, each sample

bottle was fully filled with the water sample without entrapping air bubbles. Each bottle was labeled with its corresponding sampling station and time of sampling. The collected samples were kept at 4 °C to minimize the microbial activity in the water (APHA 2005). Generally, water samples containing colloidal or suspended particulate material could interfere with the metal analysis. The samples were immediately filtered with 0.45 μm cellulose acetate membrane filter (Whatman Milipores, Clifton, NJ) after being transported to the laboratory. This procedure is crucial to prevent the occurrence of clogging during analysis with

spectrometry instruments and to obtain the dissolved ions for metal analysis (APHA 2005). Then, samples were acidified with HNO₃ to pH <2 in order to prevent precipitation of the components, such as metal oxides and hydroxides, and to retard any biological activities (APHA 2005).

Water analyses

Multiparameter probes (SevenGo pro probe and SevenGo Duo pro probe, Mettler Toledo AG, Switzerland) were used to conduct in situ measurement of electrical conductivity (EC), total dissolved solids (TDS), salinity, redox potential (Eh), and pH. The temperature and dissolved oxygen (DO) were measured using a YSI 52-dissolved oxygen meter (YSI Inc., Yellow Springs, Ohio). All probes were calibrated prior to sampling. Bicarbonate (titration method using 0.02 N HCl) and chloride ions (argentometric method using 0.0141 N AgNO₃) were analyzed on site using unfiltered samples (APHA 2005). Meanwhile, the filtered samples were separated into two polyethylene bottles. The first bottle was for subsequent analysis of sulfate (SurfaVer 4 HACH method) and nitrate (NitraVer 5 HACH method) and the second bottle was for the determination of cations (Ca, Na, Mg, and K) and metals (²⁷Al, ⁷⁵As, ¹³⁸Ba, ⁹Be, ¹¹¹Cd, ⁵⁹Co, ⁶³Cu, ⁵²Cr, ⁵⁷Fe, ⁵⁵Mn, ⁶⁰Ni, ²⁰⁸Pb, ⁸⁰Se, and ⁶⁶Zn). The cations (Ca, Na, Mg, and K) were analyzed by flame atomic absorption spectrometry (FAAS, Shimadzu AA6800) while the trace metals were analyzed by inductive couple plasma mass spectrometry (ICP-MS, ELAN DRC-e, Perkin Elmer).

Quality control and quality assurance were applied on samples and data collected in order to ensure the overall precision and accuracy of the data. Sampling, preservation, and transportation of the water samples to the laboratory were based on the Standard Method for Water and Wastewater Analysis (APHA 2005). All the reagents used were of analytical grade or equivalent and free from any contaminants. All the laboratory apparatus was pre-cleaned with 5 % (v/v) concentrated nitric acid (HNO₃) and then rinsed with distilled water (APHA 2005). This procedure is crucial to ensure that any contaminants and traces of cleaning reagent were removed before the analysis (APHA 2005). Polyethylene bottles (free from material that may contain metals) were used for collecting the water samples in order to avoid and minimize interference for heavy metal analysis (APHA 2005). The accuracy of the result was also determined by performing triplicate samples ($n=3$) with relative standard deviation. Blanks and calibration standards were used throughout the FAAS and ICP-MS analyses. Standard solutions were prepared using stock standard solutions with Milli-Q water (water resistivity >18.2 Mohms·cm at 25 °C; Millipore, MA, USA). Blanks were determined for background correction. The concentration of trace metals were expressed as micrograms per liter and milligrams per liter

for cations. The accuracy of the ICP-MS performance was assessed by external standards, which were prepared by diluting the ICP Multi-Element Mixed Standard III (Perkin Elmer) into a series of concentrations with the same acid mixture used for sample dissolution. The recoveries of trace elements ranged from 95 to 105 % (± 5 %), as shown in Table 1.

Data analyses

All statistical analyses were performed using the PASW Statistics 18 (formerly known as SPSS Statistics 18, or SPSS Base). ANOVA was applied to test the significant difference for all water quality variables among stations. A post hoc test was performed using the least significant difference test with a degree of significance at 0.05. The chemometric approach was performed through HCA and PCA (Singh et al. 2004; Praveena et al. 2011). HCA was first applied to the spatial variations among the stations, followed by the use of PCA to extract, and distinguish the potential factors or sources of pollution contributing to the variations of the water quality measures.

Hierarchical cluster analysis

In this study, HCA was used to investigate the groupings of the sampling points. This is the most common approach in which clusters are formed sequentially. This approach classifies variables or cases/observations into classes (clusters) on the basis of similarities within a class and dissimilarities between different classes from the dataset with respect to the predetermined characteristics

Table 1 Percentage recoveries of trace metals by ICP-MS

Metals	Recovery (%)	Detection limit ranges (ppt)
Al	100.28	1–10
Ba	99.01	<0.1–1
Cd	102.05	1–10
Cu	101.87	1–10
Fe	98.46	1–10
Pb	102.12	<0.1–1
Mn	97.57	1–10
Zn	101.97	1–10
Co	100.53	1–10
Be	103.08	1–10
As	102.29	1–10
Cr	101.53	1–10
Ni	101.72	1–10
Se	103.43	10–100

(Boyacioglu and Boyacioglu 2008; Praveena et al. 2011). It is a useful technique to investigate spatial and temporal variations (Singh et al. 2005; Praveena et al. 2011). HCA was performed on the river water quality data to group similar sampling points within the Langat River. The squared Euclidean was applied as a distance matrix and Ward’s method as a linkage method (Singh et al. 2005). Ward’s clustering procedure is acknowledged to be the best method (Reghunath et al. 2002). It was used for the calculation in HCA since it yields a larger proportion of correct classified observations than other methods. The result of a hierarchical clustering procedure can be displayed graphically using a tree diagram, also known as a dendrogram. A dendrogram distinguishes groups of high similarity that have small distances between clusters while the dissimilarity between groups is represented by the maximum of all possible distances between clusters. A dendrogram shows a picture of the group and their proximity with a

dramatic reduction in the dimensionality of the original data (Shrestha and Kazama 2007; Alkarkhi et al. 2009a, b). Moreover, previous studies also showed the reliability of HCA in the classification of water quality and as a guide for future sampling strategies (Singh et al. 2004; Shrestha and Kazama 2007; Alkarkhi et al. 2009a, b; Praveena et al. 2011).

Principal component analysis

Osman et al. (2012) stated that PCA is an exploratory, multivariate, statistical technique that can be used to examine data variability. It is a useful technique employed to find the optimal ways of combining variables into a small number of subsets. PCA attempts to explain the variance of a large set of intercorrelated variables by transforming them into a smaller set of independent variables and reduce the complexity of data into principal components (Singh et al. 2004, 2005). In

Table 2 Descriptive analysis for selected water matrices at Langat River (*n*=90)

	Unit	Minimum	Maximum	Range	Mean±SD	CV
pH	–	4.79	7.48	2.69	6.67±0.53	7.95
Eh	mV	–17.60	150.60	168.20	32.95±33.17	100.68
Temp	°C	26.80	29.90	3.10	28.27±1.03	3.63
EC	mS/cm	0.09	37.40	37.32	14.55±16.18	111.16
Sal	ppt	0.05	23.70	23.65	9.07±10.18	112.17
DO	mg/L	1.50	5.40	3.90	3.25±1.14	35.18
TDS	mg/L	42.40	18,720.00	18,677.60	7,287.71±8,101.91	111.17
Na	mg/L	5.21	11,183.00	11,177.79	4,022.32±4,543.16	112.95
Ca	mg/L	2.55	356.15	353.60	134.18±151.52	112.92
K	mg/L	0.56	171.95	171.39	47.68±61.56	129.11
Mg	mg/L	0.70	1,230.70	1,230.00	452.24±508.22	112.38
HCO ₃	mg/L	9.76	139.08	129.32	61.92±45.19	72.98
Cl	mg/L	10.00	8,347.41	8,337.41	3,001.39±3,409.88	113.61
NO ₃	mg/L	0.50	12.10	11.60	3.25±2.97	91.22
SO ₄	mg/L	2.00	1,325.00	1,323.00	508.78±570.22	112.08
Al	µg/L	1.15	5,191.70	5,190.56	290.07±817.50	281.83
Ba	µg/L	<0.005	30.71	30.71	7.14±7.32	102.57
Cd	µg/L	<0.01	0.53	0.53	0.11±0.12	103.79
Cu	µg/L	0.57	81.02	80.45	28.16±29.70	105.49
Fe	µg/L	81.79	807.15	725.36	312.60±156.74	50.14
Pb	µg/L	<0.005	6.99	6.99	1.07±1.64	153.25
Mn	µg/L	<0.005	640.07	640.07	87.14±119.60	137.26
Zn	µg/L	0.95	65.06	64.11	15.21±14.29	93.94
Co	µg/L	0.06	6.22	6.17	0.64±1.14	180.15
Be	µg/L	<0.05	1.10	1.10	0.13±0.23	176.29
As	µg/L	0.08	24.71	24.63	8.54±9.15	107.03
Cr	µg/L	0.32	4.67	4.35	1.13±0.91	80.03
Ni	µg/L	0.80	24.72	23.92	7.29±6.75	92.56
Se	µg/L	<0.10	265.00	265.00	87.86±101.31	115.31

NA not available, *Eh* redox potential, *Temp* temperature, *EC* electrical conductivity, *Sal* salinity, *DO* dissolved oxygen, *TDS* total dissolved solids, *SD* standard deviation, *CV* coefficient of variance

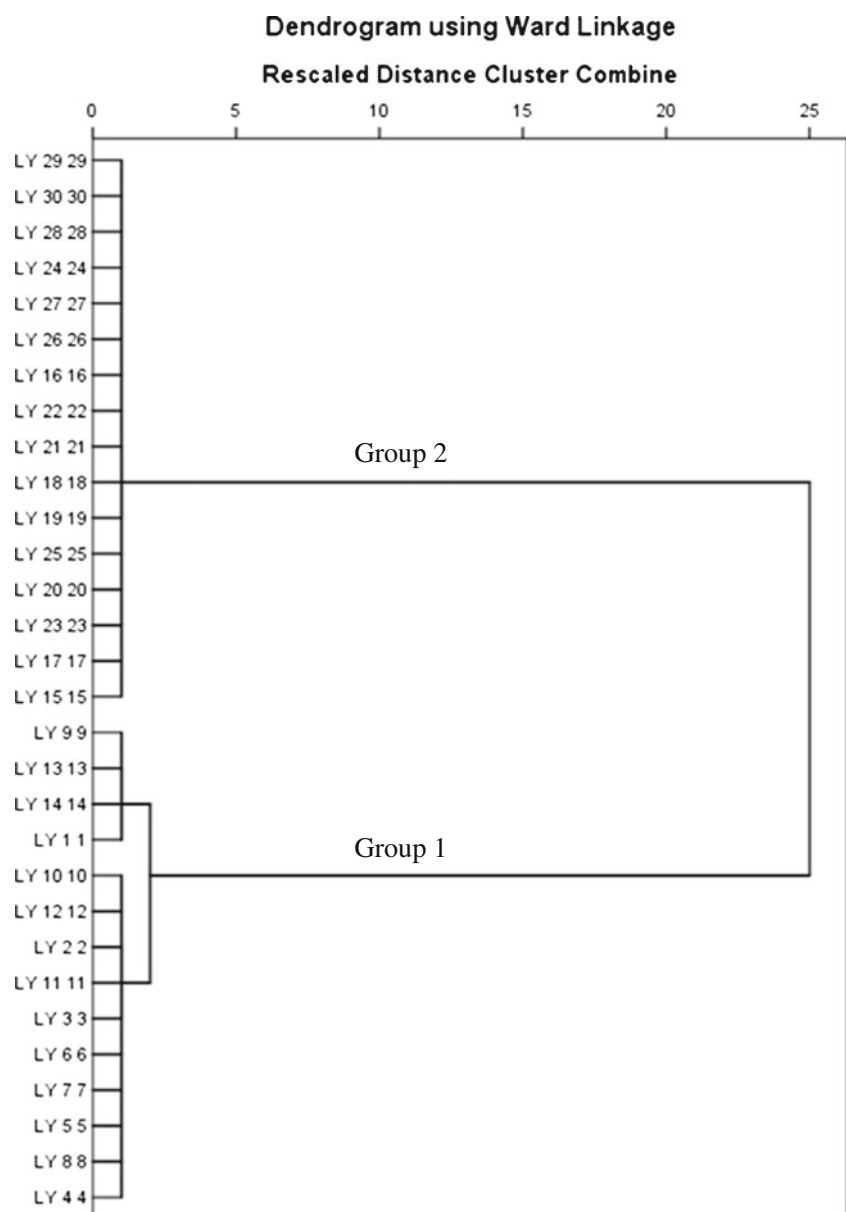
this study, PCA was applied in datasets that had been pre-clustered by extracting the eigenvalues and eigenvectors from a square matrix produced by multiplying data matrix. The most significant components were extracted to reduce the contribution of variables with minimum significance. Then, the obtained components were further subjected to varimax rotation to generate varimax factors and maximize the differences between the variables, thus facilitating easy interpretation of the data. A principal component provides information on the most meaningful parameters, which describes a whole dataset, affording data reduction with a minimum loss of the original information (Shrestha and Kazama 2007). The components are ordered in such a way that the first PC explains most of the variance in the data, and each subsequent one accounts for the largest proportion of variability that has not

been accounted for by its predecessors. This is to clearly differentiate potential factors or pollution sources contributing to the variation of water quality.

Results and discussion

Table 2 shows the descriptive statistics for the selected physicochemical parameters, major ions, and trace metal concentrations. The coefficients of variance (CV) for all variables were above 50 % except for the pH and temperature. The CV was calculated based on the sum value of standard deviation from each studied metal divided by its mean value. The high CV indicated a high variation between sampling stations. In addition, one-way ANOVA analysis

Fig. 2 Dendrogram showing hierarchical cluster analysis between stations



also proved that the studied variables varied significantly among the stations ($p < 0.05$).

Hierarchical cluster analysis

In this study, HCA was applied to detect similarities between the sampling stations. A total of 29 variables which included physicochemical parameters (temperature, EC, TDS, salinity, DO, pH, and Eh), major ions (HCO_3 , Cl, SO_4 , NO_3 , Ca, Na, K, and Mg), and trace metals (Al, Ba, Be, Cd, Co, Cu, Cr, Fe, Mn, Ni, Pb, Se, and Zn) were first subjected to HCA. The dendrogram of the locations of different sites along the study area applied for water datasets are presented in Fig. 2. It shows that the 30 sampling stations can be grouped into two clusters (namely group 1 and group 2). The results indicate the potential contributing sources, which are attributed to both natural and anthropogenic origin. Group 1 accounts for sampling stations LY 1 to LY 14, which are mainly located in the vicinity of agricultural land and the Strait of Malacca (Fig. 1). The movement of seawater during the tidal flow has significantly contributed to the high load of salinity, EC, TDS, and also additional ions notably K, Mg, and Na within downstream of the Langat River. Thus, LY 1 to LY 14 were grouped under the same cluster and can be denoted as a group being governed by seawater intrusion. Group 2 consisted of sampling stations LY 15 to LY 30, which are mainly located in the eastern part of the study area. This area has experienced urbanization and the land use pattern is predominantly that of urban activities and agricultural fields (DOA 1995; JICA and MGD 2002; Juahir et al. 2011; Osman et al. 2012; Fig. 1). The sampling

stations are mainly located further inland from the estuary and in close proximity to the major pollution sources, such as industrial and domestic discharge. As such, sampling stations from group 2 receive minimal impact from seawater intrusion compared to group 1. In addition, the differences between group 1 and group 2 can likewise be substantiated by the changes in water type from Na-Cl facies (LY 1 to LY 14) to Ca- HCO_3 facies (LY 15 to LY 30) (downstream to upstream), as depicted in Fig. 3.

Principal component analysis

The PCA was applied on the water quality dataset to identify the spatial sources of pollution within group 1 and group 2 in the Langat River (Tables 3 and 4). In reference to the eigenvalues (greater than 1), four components were extracted in group 1 and explained 85 % of the total variance (Table 3), whereas group 2 extracted six components with a total variance of 88 % (Table 4). Comparable loadings were observed in group 1 and group 2.

In Table 3 (group 1), PC 1 accounted for 35 % of the total variance. This component showed high loading of EC, salinity, TDS, Ca, HCO_3 , SO_4 , Na, Mg, temperature, DO, Eh, pH, K, NO_3 , and As (Table 3). EC, salinity, and TDS are commonly regarded as indicator for the presence of dissolved ions including inorganic salt and organic matter in water (Reza and Singh 2010). Generally, the EC and TDS increases as the dissolved ions increase. Similarly, as the dissolved salt concentration increases, the salinity will also increase (Connell and Miller 1984; Elder 1988). Such a statement is also supported by the strong positive component loadings for EC, salinity, and TDS in PC 1 (Table 3). Meanwhile, the high loadings for the major ions (Ca, Na,

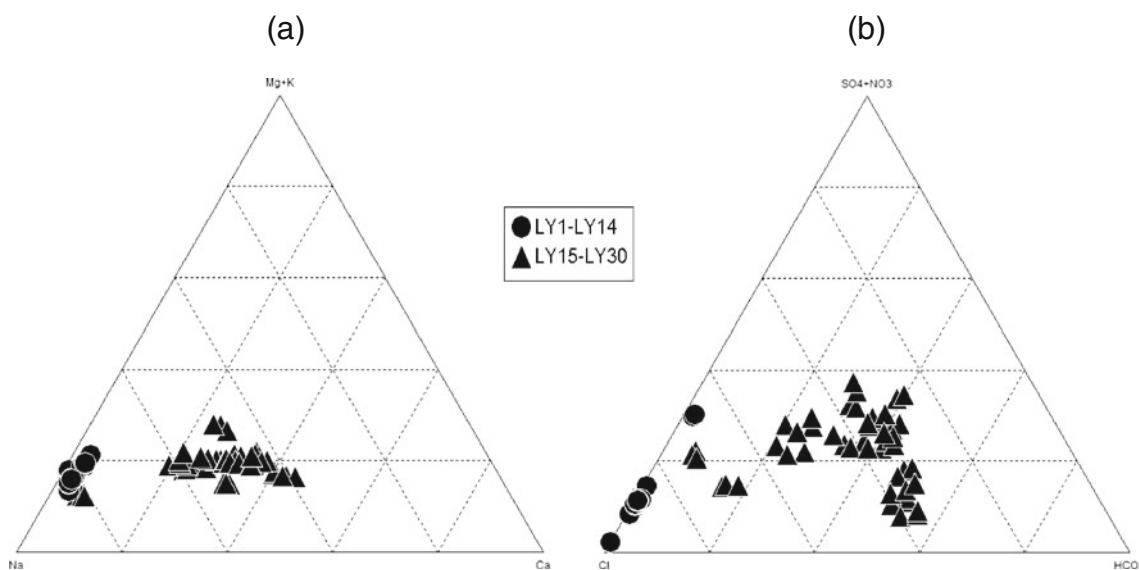


Fig. 3 Ternary plots for a cations and b anions of water samples in Langat River

Table 3 Component loadings of river water quality variables on varimax rotated matrix for group 1

	Component			
	PC 1	PC 2	PC 3	PC 4
Electrical conductivity	<i>0.94</i>	-0.04	0.21	0.18
Salinity	<i>0.94</i>	-0.04	0.21	0.19
Total dissolved solids	<i>0.94</i>	-0.05	0.21	0.19
Ca	<i>0.93</i>	-0.33	0.06	0.05
HCO ₃	<i>0.83</i>	-0.46	-0.02	-0.07
SO ₄	<i>0.80</i>	-0.50	-0.22	-0.04
Na	<i>0.79</i>	-0.03	0.08	0.40
Mg	<i>0.75</i>	-0.22	0.44	0.40
Temperature	<i>0.72</i>	0.39	0.29	-0.03
Dissolved oxygen	<i>0.72</i>	0.37	0.15	0.35
Eh	-0.72	0.62	0.19	0.17
pH	<i>0.72</i>	-0.62	-0.19	-0.17
K	<i>0.69</i>	-0.34	0.45	0.34
NO ₃	-0.65	-0.07	-0.12	-0.31
As	<i>0.63</i>	0.41	0.54	0.31
Co	-0.12	<i>0.98</i>	-0.05	-0.10
Be	-0.17	<i>0.96</i>	0.03	-0.12
Al	-0.25	<i>0.95</i>	0.11	0.05
Mn	-0.13	<i>0.95</i>	-0.13	-0.18
Ba	-0.16	<i>0.94</i>	-0.17	-0.21
Ni	0.27	<i>0.90</i>	0.14	0.21
Zn	-0.26	<i>0.86</i>	-0.07	-0.11
Fe	0.34	<i>0.79</i>	-0.06	-0.15
Pb	0.12	0.43	-0.78	0.04
Cl	0.15	-0.13	<i>0.69</i>	0.64
Cd	0.16	0.05	<i>0.57</i>	-0.08
Cr	0.36	0.20	<i>0.55</i>	0.11
Se	0.18	-0.09	-0.11	<i>0.68</i>
Cu	0.40	-0.28	0.18	<i>0.62</i>
Initial eigenvalue	10.21	9.04	2.93	2.38
Percent of variance	35.22	31.17	10.09	8.21
Cumulative percent	35.22	66.39	76.48	84.68

The values in italics are factor loadings above 0.50 that taken after varimax rotation was performed

Mg, K, HCO₃, SO₄, and NO₃) in PC 1 may be explained by the mixing condition between freshwater (river) and seawater (Aris et al. 2007; Praveena et al. 2011; Aris et al. 2012). In addition, forest and agriculture are regarded as the primary landuse found in the sampling locations and were included within group 1 (DOA 1995; Juahir et al. 2011; Fig. 1). Taking this into consideration, the observed high loading of As and NO₃ in PC 1 also implies the possible contribution from agricultural land. Furthermore, farm draining during the rainy season led pollution of the river caused by the by-products from the agricultural applications

(Diagomanolin et al. 2004; Shokrzadeh and Saeedi Saravi 2009). Such relatively high loadings strongly indicate that the river water in group 1 is primarily controlled by seawater intrusion and agricultural discharges. PC 2, with a total variance of 23.17 %, consists of Co, Be, Al, Mn, Ba, Ni, Zn, and Fe (Table 3). This component was primarily contributed by trace metals. The elements especially Al, Mn, Ni, Zn, and Fe have been constantly released into the environment through weathering processes (Alloway 1995). The high concentrations of Al, Fe, and Mn in the Langat River are due to the composition of sediments, which is predominantly controlled by its lithology. Ferralsols (oxisols and ultisols), which are rich in Al and Fe, are acidic and highly weathered (Alloway 1995). During the rainy season, intense rainfall erodes the topsoil and carries these elements into the river. In this case, this component can be attributed to rock weathering. Meanwhile, the Fe, Al, and Mn oxides have a profound effect in controlling the adsorption and flocculation of other elements in the sediment (Alloway 1995). PC 3, with a total variance of 10 %, consists of Pb, Cl, Cd, and Cr (Table 3). Negative loadings of Pb and positive loading of Cl, Cd, and Cr suggest the occurrence of ion competition between these element binding sites (Campbel and Stokes 1985). Furthermore, Abdullah and Royle (1974) observed that an increase in salinity caused a decrease in concentrations of Zn, Cu, Fe and Cd, whereas a reverse trend was observed in for Pb and Zn. PC 4 with a total variance of 8 % consists of Se and Cu (Table 3). This component was deemed to be attributed to the pig farming activities (UPUM 2002; Lee et al. 2006). Copper sulfate is normally added to the animal feed as an additive to control certain diseases (Sarmani et al. 1992). The by-products from pig food contributed to the elevated Cu to the river via the effluent discharged (Sarmani et al. 1992; UPUM 2002; Lee et al. 2006; Juahir et al. 2011).

Group 2 comprised sampling stations LY 15 to LY 30. In Table 4 (group 2), PC 1, with a total variance of 35 %, consists of salinity, EC, TDS, Mg, Na, Se, Cu, Cl, Ni, SO₄, and Cr. The component loadings are comparable with PC 1 in group 1. The similar loading in EC, salinity, TDS, and major ions indicates that the seawater intrusion still influence the hydrochemistry of the study area. Meanwhile, the high loading of Se and Cu in PC 1 may be attributed to the extraction of selenium dioxide from residues obtained during the purification of copper (Langner 2000; Hait et al. 2009). In addition, the fluctuation in flow between freshwater and seawater causes elevated salt concentration, which, consequently, increases the competition between the cations and the trace metals for binding sites in the particulates (Connell and Miller 1984; Elder 1988). The cations, being more prominent, drive the trace metals into the overlying water column. As a result, metals may be desorbed from the sediment thereby increasing their concentrations (Connell and Miller 1984; Elder 1988). PC 2, with a total variance of

Table 4 Component loadings of river water quality variables on varimax rotated matrix for group 2

	Component					
	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Salinity	<i>0.98</i>	0.14	0.05	0.00	0.04	0.00
Electrical conductivity	<i>0.97</i>	0.14	0.07	0.04	0.09	0.04
Total dissolved solids	<i>0.97</i>	0.14	0.07	0.04	0.09	0.04
Mg	<i>0.96</i>	0.15	0.20	−0.01	0.05	0.05
Na	<i>0.95</i>	0.18	−0.08	0.04	0.10	0.04
Se	<i>0.84</i>	0.24	0.36	−0.01	0.10	0.22
Cu	<i>0.81</i>	0.44	0.06	0.13	−0.18	−0.08
Cl	<i>0.81</i>	−0.07	0.31	−0.18	−0.06	−0.18
Ni	<i>0.80</i>	0.36	0.35	0.12	0.04	−0.14
SO ₄	<i>0.70</i>	0.22	0.41	−0.41	0.01	0.17
Cr	<i>0.66</i>	0.62	0.05	0.10	−0.16	−0.11
Pb	−0.04	<i>0.92</i>	0.05	0.02	−0.19	−0.04
Be	0.18	<i>0.91</i>	0.19	−0.08	−0.09	−0.01
Zn	0.36	<i>0.86</i>	−0.16	0.05	−0.13	0.03
Fe	−0.06	<i>0.85</i>	0.18	0.20	0.11	−0.03
Co	0.22	<i>0.84</i>	0.28	−0.03	0.16	0.01
Al	0.40	<i>0.82</i>	−0.06	0.15	−0.09	−0.01
As	0.36	<i>0.79</i>	0.38	−0.04	0.05	0.00
Mn	0.22	<i>0.65</i>	0.28	−0.08	0.39	0.24
HCO ₃	−0.52	− <i>0.55</i>	0.22	−0.12	0.16	−0.11
pH	−0.39	−0.10	− <i>0.84</i>	−0.02	0.00	0.06
Eh	0.46	0.09	<i>0.83</i>	0.00	0.03	−0.07
K	0.26	−0.25	− <i>0.80</i>	0.20	0.14	−0.16
Temperature	0.62	0.16	<i>0.64</i>	−0.07	0.17	0.21
Ba	−0.14	0.38	0.14	<i>0.72</i>	0.27	0.06
Dissolved oxygen	0.03	−0.10	−0.14	<i>0.70</i>	−0.10	−0.10
Cd	0.47	0.48	−0.15	<i>0.58</i>	−0.15	0.15
Ca	0.12	−0.14	−0.06	0.00	<i>0.94</i>	0.02
NO ₃	0.03	−0.03	0.06	−0.05	0.03	<i>0.97</i>
Initial eigenvalue	10.16	7.31	3.55	1.73	1.43	1.29
Percent of variance	35.04	25.20	12.24	5.96	4.95	4.43
Cumulative percent	35.04	60.24	72.48	78.44	83.39	87.82

The values in italics are factor loadings above 0.50 that taken after varimax rotation was performed

25 %, consists of Pb, Be, Zn, Fe, Co, Al, As, Mn, and HCO₃ (Table 4). The metals, including Pb, Be, Zn, Fe, Co, and Al, may be attributed to industrial activity, which is deemed to be closely related to the steelmaking industries (Shazili et al. 2006). In fact, the largest steelmaking industry in Malaysia is located in proximity to the upstream area (Sarmani 1989; Mokhtar et al. 2009) thus explaining the high loadings of these elements in PC 2. The high loading of Pb is related to the heavy shipping traffic and antifouling paints used (Goh and Chou 1997; Shazili et al. 2006; Berandah et al. 2010). Furthermore, intensive dredging, reclamation, construction, and shipping activities which disturb the river currents will lead to re-suspension of the sediment-bound trace elements in the environment, and, probably,

in the soluble forms readily absorbed by aquatic organisms (Zulkifli et al. 2010). PC 3, which accounted for 12 % of total variance, consists of pH, Eh, K, and temperature (Table 4). This component illustrates the influence of pH, Eh, and temperature on the quality of the river water. PC 4, PC 5, and PC 6, which explain about 6, 5, and 4 % of the total variance, respectively, have a strong positive loading on temperature as well as Ba, DO, Cd, Ca, and NO₃. The presence of Ba and Cd may be attributed to industrial activity, which is deemed to be closely related to metal finishing processes such as electroplating, etching, and preparation of metal components (Shazili et al. 2006). The NO₃ is possibly derived from geologic deposits, organic matter decomposition, untreated wastewater input, agricultural runoff, and atmospheric input (Alkarkhi et al. 2009a).

Conclusion

The present study has applied the chemometric approach to investigate the spatial variation and identify the pollution sources in the Langat River, Malaysia. The HCA rendered the sampling stations into two clusters. Cluster 1 (LY 1 to LY 14) was heavily affected by seawater while cluster 2 (LY 15 to LY 30) mainly corresponded to the agricultural and industrial activities. The cluster results suggested that certain stations should be given a high priority if remediation efforts are to be undertaken. PCA identified several intrinsic factors responsible for river pollution, either from natural or anthropogenic inputs. Group 1 extracted four components with a total variance of 85 %, while group 2 extracted six components with a total variance of 88 %. The results suggested that seawater intrusion, agricultural pollution, industrial pollution, and geological weathering were potential pollution sources for both groups. In conclusion, this study highlights the usefulness of chemometric approach in delineating factors that govern the spatial variability of hydrochemistry in a tropical river. It is evident that the chemometric approach is useful in providing a reliable classification on the basis of pollution status and identification of pollution sources. Such effort provides holistic information for effective river basin management and makes it possible to design a future spatial sampling strategy in an optimal manner.

Acknowledgments This research was funded by the Research University Grant Scheme vot no. 9199751, no. project 03-01-11-1142RU from Universiti Putra Malaysia and the Academy of Sciences for the Developing World (TWAS) project number 09-09 RG/EAS/AS_C/UNESCO FR:3240231216. The first author sincerely acknowledges the support from Graduate Research Fellowship Scholarship awarded by Universiti Putra Malaysia for her study. Rainfall data and field services provided by the Department of Drainage and Irrigation (JPS), Malaysia is greatly appreciated.

References

- Abdullah MI, Royle LG (1974) A study of the dissolved and particulate trace elements in the Bristol Channel. *J Mar Biol Assoc UK* 54(3):581–597. doi:10.1017/S0025315400022761
- Akbal F, Gürel L, Bahadır T et al (2011) Multivariate statistical techniques for the assessment of surface water quality at the mid-Black Sea Coast of Turkey. *Water Air Soil Poll* 216(1):21–37. doi:10.1007/s11270-010-0511-0
- Alkarkhi A, Ahmad A, Easa A (2009a) Assessment of surface water quality of selected estuaries of Malaysia: multivariate statistical techniques. *Environmentalist* 29(3):255–262. doi:10.1007/s10669-008-9190-4
- Alkarkhi A, Ismail N, Ahmed A, Easa A (2009b) Analysis of heavy metal concentrations in sediments of selected estuaries of Malaysia—a statistical assessment. *Environ Monit Assess* 153(1):179–185. doi:10.1007/s10661-008-0347-x
- Alloway BJ (1995) Heavy metals in soils. Blackie Academic & Professional, London
- Altın A, Filiz Z, Iscen CF (2009) Assessment of seasonal variations of surface water quality characteristics for Porsuk Stream. *Environ Monit Assess* 158(1–4):51–65. doi:10.1007/s10661-008-0564-3
- APHA (2005) Standard methods for the examination of water and wastewater, 21st edn. American Water Works Association, Water Environment Federation, Washington
- Aris AZ, Abdullah MH, Ahmed A, Woong KK (2007) Controlling factors of groundwater hydrochemistry in a small island's aquifer. *Int J Environ Sci Technol* 4(4):441–450
- Aris AZ, Praveena SM, Abdullah MH, Radojevic M (2012) Statistical approaches and hydrochemical modelling of groundwater system in small tropical island. *J Hydroinform* 14(1):206–220. doi:10.2166/hydro.2011.072
- Berandah FE, Yap CK, Ismail A (2010) Bioaccumulation and distribution of heavy metals (Cd, Cu, Fe, Ni, Pb and Zn) in the different tissues of *Chicoreus capucinus* Lamarck (Mollusca: Muricidae) collected from Sungai Janggut, Kuala Langat, Malaysia. *Environ Asia* 3(1):65–71
- Boyacioglu H, Boyacioglu H (2008) Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey. *Environ Geol* 54(2):275–282. doi:10.1007/s00254-007-0815-6
- Bu H, Tan X, Li S, Zhang Q (2010) Water quality assessment of the Jinshui River (China) using multivariate statistical techniques. *Environ Earth Sci* 60(8):1631–1639. doi:10.1007/s12665-009-0297-9
- Campbel PGC, Stokes PM (1985) Acidification and toxicity of metals to aquatic biota. *Can J Fish Aquat Sci* 42(12):2034–2049. doi:10.1139/f85-251
- Connell DW, Miller GJ (1984) Chemistry and ecotoxicology of pollution. Wiley, New York
- Davis JC (1986) Statistics and data analysis in geology, 2nd edn. Wiley, New York
- Debels P, Figueroa R, Urrutia R et al (2005) Evaluation of water quality in the Chillán River (Central Chile) using physicochemical parameters and a modified water quality index. *Environ Monit Assess* 110(1–3):301–322. doi:10.1007/s10661-005-8064-1
- Diagomanolin V, Farhang M, Ghazi-Khansari M et al (2004) Heavy metals (Ni, Cr, Cu) in the Karoon waterway river, Iran. *Toxicol Lett* 151(1):63–67. doi:10.1016/j.toxlet.2004.02.018
- DOA (1995) Landuse of Selangor and Negeri Sembilan. Department of Agriculture, Kuala Lumpur
- Elder JF (1988) Metal biogeochemistry in surface-water systems: a review of principles and concepts. In: U.S. Geological Survey circular, Issue 1013. Department of the Interior, U.S. Geological Survey, Washington, D.C.
- Gobbett DJ, Hutchison CS (1973) Geology of the Malay Peninsula: West Malaysia and Singapore. Wiley-Interscience, New York
- Goh BPL, Chou LM (1997) Heavy metal levels in marine sediments of Singapore. *Environ Monit Assess* 44(1):67–80. doi:10.1023/a:1005763918958
- Hait J, Jana RK, Sanyal SK (2009) Processing of copper electrorefining anode slime: a review. *Miner Process Extract Metall Rev* 118(4):240–252. doi:10.1179/174328509X431463
- Huang J, Ho M, Du P (2011) Assessment of temporal and spatial variation of coastal water quality and source identification along Macau peninsula. *Stoch Environ Res Risk Assess* 25(3):353–361. doi:10.1007/s00477-010-0373-4
- Hussain M, Ahmed SM, Abderrahman W (2008) Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia. *J Environ Manag* 86(1):297–307. doi:10.1016/j.jenvman.2006.12.007
- JICA and MGDM (2002) The study on the sustainable groundwater resources and environmental management for the Langat Basin in Malaysia. Final Report. vol 3. Kuala Lumpur, Malaysia
- Juahir H, Zain S, Yusoff M et al (2011) Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environ Monit Assess* 173(1):625–641. doi:10.1007/s10661-010-1411-x

- Kazi TG, Arain MB, Jamali MK et al (2009) Assessment of water quality of polluted lake using multivariate statistical techniques: a case study. *Ecotoxicol Environ Saf* 72(2):301–309. doi:10.1016/j.ecoenv.2008.02.024
- Krishna AK, Satyanarayanan M, Govil PK (2009) Assessment of heavy metal pollution in water using multivariate statistical techniques in an industrial area: a case study from Patancheru, Medak District, Andhra Pradesh, India. *J Hazard Mater* 167(1–3):366–373. doi:10.1016/j.jhazmat.2008.12.131
- Langner BE (2000) Selenium and selenium compounds. In: Ullmann's encyclopedia of industrial chemistry. Wiley, Weinheim. doi:10.1002/14356007.a23_525
- Lee YH, Abdullah MP, Chai SY, Mokhtar MB, Ahmad R (2006) Development of possible indicators for sewage pollution for the assessment of Langat River ecosystem health. *Malays J Anal Sci* 10(1):15–26
- Mokhtar MB, Aris AZ, Munusamy V, Praveena SM (2009) Assessment level of heavy metals in *Penaeus monodon* and *Oreochromis* spp. in selected aquaculture ponds of high densities development area. *Eur J Sci Res* 30(3):348–360
- Najar I, Khan A (2012) Assessment of water quality and identification of pollution sources of three lakes in Kashmir, India, using multivariate analysis. *Environ Earth Sci* 66(8):2367–2378. doi:10.1007/s12665-011-1458-1
- Osman R, Saim N, Juahir H, Abdullah M (2012) Chemometric application in identifying sources of organic contaminants in Langat river basin. *Environ Monit Assess* 184(2):1001–1014. doi:10.1007/s10661-011-2016-8
- Praveena SM, Abdullah MH, Bidin K, Aris AZ (2011) Understanding of groundwater salinity using statistical modeling in a small tropical island, East Malaysia. *Environmentalist* 31(3):279–287. doi:10.1007/s10669-011-9332-y
- Praveena SM, Ahmed A, Radojevic M et al (2008) Multivariate and geoaccumulation index evaluation in mangrove surface sediment of Mengkabong Lagoon, Sabah. *Bull Environ Contam Toxicol* 81(1):52–56. doi:10.1007/s00128-008-9460-3
- Reghunath R, Murthy TRS, Raghavan BR (2002) The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. *Water Res* 36(10):2437–2442. doi:10.1016/S0043-1354(01)00490-0
- Reza R, Singh G (2010) Heavy metal contamination and its indexing approach for river water. *Int J Environ Sci Technol* 7(4):785–792
- Saim N, Osman R, Sari AbgSpian DR et al (2009) Chemometric approach to validating faecal sterols as source tracer for faecal contamination in water. *Water Res* 43(20):5023–5030. doi:10.1016/j.watres.2009.08.052
- Sarmani S (1989) The determination of heavy metals in water, suspended materials and sediments from Langat River, Malaysia. *Hydrobiologia* 176–177(1):233–238. doi:10.1007/bf00026558
- Sarmani S, Abdullah MP, Baba I, Majid AA (1992) Inventory of heavy metals and organic micropollutants in an urban water catchment drainage basin. *Hydrobiologia* 235–236(1):669–674. doi:10.1007/bf00026255
- Shazili NAM, Yunus K, Ahmad AS et al (2006) Heavy metal pollution status in the Malaysian aquatic environment. *Aquat Ecosyst Health Manag* 9(2):137–145. doi:10.1080/14634980600724023
- Shokrzadeh M, Saeedi Saravi SS (2009) The study of heavy metals (zinc, lead, cadmium, and chromium) in water sampled from Gorgan coast (Iran), Spring 2008. *Toxicol Environ Chem* 91(3):405–407. doi:10.1080/02772240902830755
- Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. *Environ Model Softw* 22(4):464–475. doi:10.1016/j.envsoft.2006.02.001
- Simeonov V, Tsakovski S, Lavric T et al (2004) Multivariate statistical assessment of air quality: a case study. *Microchim Acta* 148(3–4):293–298. doi:10.1007/s00604-004-0279-2
- Singh KP, 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. *Water Res* 38(18):3980–3992. doi:10.1016/j.watres.2004.06.011
- Singh KP, Malik A, Singh VK et al (2005) Chemometric analysis of groundwater quality data of alluvial aquifer of Gangetic plain, North India. *Anal Chimica Acta* 550(1–2):82–91. doi:10.1016/j.aca.2005.06.056
- Taha M (2003) Groundwater and geoenvironmental quality issues in the Langat Basin, Malaysia. In: Kono I, Nishigaki M, Komatsu M (eds) *Groundwater engineering—recent advances: Proceedings of the International Symposium, Okayama, Japan, 2003*. Taylor & Francis
- UPUM (2002) Program Pencegahan dan Peningkatan Kualiti Air Sungai Langat. Final Draft Report. Universiti Malaya Consultancy Unit, Kuala Lumpur, Malaysia
- Varol M, Gökot B, Bekleyen A, Şenc B (2012) Spatial and temporal variations in surface water quality of the dam reservoirs in the Tigris River basin, Turkey. *Catena* 92:11–21. doi:10.1016/j.catena.2011.11.013
- Zulkifli S, Mohamat-Yusuff F, Arai T et al (2010) An assessment of selected trace elements in intertidal surface sediments collected from the Peninsular Malaysia. *Environ Monit Assess* 169(1):457–472. doi:10.1007/s10661-009-1189-x