Environ Monit Assess (2012) 184:1001–1014
DOI 10.1007/s10661-011-2016-8

Chemometric application in identifying sources of organic contaminants in Langat river basin

Rozita Osman · Norashikin Saim · Hafizan Juahir · Md Pauzi Abdullah

Received: 8 December 2009 / Accepted: 16 March 2011 / Published online: 15 April 2011

Abstract Increasing urbanization and changes in land use in Langat river basin lead to adverse impacts on the environment compartment. One of the major challenges is in identifying sources of organic contaminants. This study presented the application of selected chemometric techniques: cluster analysis (CA), discriminant analysis (DA), and principal component analysis (PCA) to classify the pollution sources in Langat river basin based on the analysis of water and sediment samples collected from 24 stations, monitored for 14 organic contaminants from polycyclic aromatic hydrocarbons (PAHs), sterols, and pesticides groups. The CA and DA enabled to group 24 monitoring sites into three groups of pollution source (industry and urban socioeconomic, agricultural activity, and urban/domestic sewage) with five major discriminating variables: naphthalene, pyrene, benzo[a]pyrene, coprostanol, and cholesterol. PCA analysis, applied to water data sets, resulted in four latent factors explaining 79.0% of the total variance while sediment samples gave five latent factors with 77.6% explained variance. The varifactors (VFs) obtained from PCA indicated that sterols (coprostanol, cholesterol, stigmasterol, β-sitosterol, and stigmastanol) are strongly correlated to domestic and urban sewage, PAHs (naphthalene, acenaphthene, pyrene, benzo[a]anthracene, and benzo[a]pyrene) from industrial and urban activities and chlorpyrifos correlated to samples nearby agricultural sites. The results demonstrated that chemometric techniques can be used for rapid assessment of water and sediment contaminations.

Keywords Chemometric · Organic contaminants · Cluster analysis · Discriminant analysis · Principal component analysis

R. Osman (✉) · N. Saim
Faculty of Applied Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
e-mail: rozit471@salam.uitm.edu.my

H. Juahir
Faculty of Environmental Study, Universiti Putra Malaysia, 43000 Serdang, Selangor, Malaysia

Md. P. Abdullah
Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Introduction

Movement of organic contaminants into water and soil take place through disposals of waste, effluent discharges from industries and chemicals release through agricultural activities. Currently, in Malaysia, the water resource issues have grown in

magnitude and complexity compared to 20 years ago. This can be attributed to the shift of the Malaysian economy from agriculture in the 1970s to industry-based in the 1990s (Juahir 2008). For the past three decades, the Langat river basin is experiencing rapid land use change drive by the development process. In a move to identify sources of river water pollution, a total of 120 river basins was monitored with 926 monitoring stations established in 2004 (DOE 2006). The Department of Environment (DOE), Malaysia, established a network of sampling stations along Langat river basin for determination of some water quality criteria for monitoring purposes. Based on these monitoring data, sewage treatment plants, manufacturing sector, livestock farming, and agro-based industries were identified as the major sources of water pollution in Malaysia. The Langat river basin is one of the most studied river basins in Malaysia. Study conducted by Mohamed et al. (2009) reported that agricultural and industrial activities were identified as the main pollution sources to groundwater and soil ecosystem in the Langat basin. These activities contributed towards the generation of non-point sources (NPS) of pollutants. However, the source of pollutants produced were difficult to be exactly identified. Most of the studies focus on the water quality index to evaluate the status of river pollution. The objective of this study is to identify the sources of organic contaminants in environment compartments using chemometrics approach. Several organic contaminants from different class; polycyclic aromatic hydrocarbons (PAHs), pesticides (chlorpyrifos and cypermethrin), and sterols were selected as parameters to represent their sources. Most of the organic contaminants discharged into waters rapidly become associated with particulate matter and incorporated in sediments. So, both water and sediment samples from this river basin will be analysed. The data sets obtained were subjected to chemometric techniques such as cluster analysis (CA), discriminant analysis (DA), and principal component analysis (PCA) in order to investigate the compositional differences between water and sediment samples, and to identify the pollution sources. Since there are too many point sources (PS) and NPS pollution along the Langat river, it is quite a challenge to identify the origin of each pollutant observed in the water.

Chemometric techniques have often been used in exploratory data analysis tools for classification (Brodnjak-Voncina et al. 2002; Kowalkowski et al. 2006) of samples (observations) or sampling stations and identification of pollution sources (Shrestha and Kazama 2007; Vega et al. 1998). This is a useful technique for identifying common patterns in data distribution that allow the identification of possible factors/sources, which influence water systems as well as a rapid solution to pollution problems (Simeonov et al. 2003). In many other cases, the exploratory data analysis results will serve to achieve an insight into, e.g., the contamination situation of a certain location and to make a plan for remediation or to prepare more focused sampling plans. Principal component analysis is an exploratory, multivariate, statistical technique that can be used to examine data variability. The principal 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. Although the number of PCs equals the number of independent original variables, generally, most of the variation in the data sets can be explained by the first few principal components that can be used to represent the original observations (Abdul-Wahab et al. 2005). Multivariate techniques can consider a number of factors which control data variability simultaneously (Boruvka et al. 2005) and therefore, offer significant advantages over univariate techniques. Recently, self-organizing maps technique is used to give a more specific classification (Astel et al. 2007). Aquatic environment such as river basin mainly affected by pollution from point and non-point sources. Identifiable pollutant input is discharged via a drain (but not exclusively) from industrial and municipal treatment plants are classified as PS pollutants. In contrast, NPS pollution enters the water in a distributed, cumulative way. Urban storm water, constitutes the primary transport mechanism that introduces NPS. To identify the origin of each pollutant observed in the water and sediment is a challenge task as there is too many PS and NPS pollution along Langat river.

Experimental

Sampling locations

Sampling stations were selected based on Department of Environment sampling stations representing water outlet for domestic, agricultural, and industrial activities. Sampling of water and sediment samples along Langat river basin was conducted from 4th June to 7th June 2008 involved 24 sampling stations. In water sampling, the grab sample technique was used (vertical grab sampler 5 L, Ocean Test Equipment, Florida, USA). The water was fished out of the river and transferred into 2.5 L amber bottles. Water samples were collected in duplicates and were acidified to pH 2 with sulfuric acid to eliminate biological activity in the water. Prior to extraction, 1,000 mL of water samples were filtered through a 0.45-μm glass fiber filter (Whatman International Ltd Maidstone, England) to remove suspended matter. Methanol (10%) was added prior to solid phase extraction (SPE).

Chemical and reagents

All solvents (methanol, dichloromethane, *n*-hexane, acetone,) were of GC grade or higher and purchased from Merck (Darmstadt, Germany). Silica gel (70–230 mesh ASTM) was obtained from Merck (Darmstadt, Germany) and diatomaceous earth non-washed was bought from Sigma-Aldrich (Steinheim, Germany). Silica gel was activated for 24 h at 130°C before used. Chlorpyrifos PESTANAL® 99.5%, cypermethrin mix of isomers PESTANAL® 98%, and *N*, *O*-bis(trimethylsilyl)trifluoroacetamide + trimethylchlorosilane, 99:1 were purchased from Sigma-Aldrich. Individual standards of PAHs: naphthalene, acenaphthene, antracene, and pyrene were obtained from Dr. Ehrenstorfer, GmbH (Augsburg, Germany); acenaphthalene, fluorene, benzo[a]anthracene, and benzo[a]pyrene were obtained from Supelco (Bellefonte, USA). Individual standards of sterols: 5β-cholestan-3β-ol (coprostanol), 5-cholesten-3β-ol (cholesterol), 5β-cholestan-3α-ol (stigmasterol), and stigmastanol were purchased from Sigma Aldrich (Steinheim, Germany). Internal standards, phenantrene d_{10} (Supelco, Bellefonte, USA) and 5α-cholestane (Sigma Aldrich, Steinheim, Germany), were used for quantification.

Solid-phase extraction

The SPE procedure was followed the method reported by Osman et al. (2009). The C_{18} cartridges (1,000 mg) were activated using 10 mL of methanol followed with 6 mL of deionized water without applying vacuum. Then, water sample (1 l) was percolated to the column using vacuum manifold at a flow rate of 6 mL min^{-1}. During the process, column should not be allowed to dry. After all samples have been loaded into the column, the column was vacuum dried for 30 min. Interference matrix was removed by eluting the column with 10 mL deionized water and again the cartridge was vacuum dried. The compound of interest was eluted by gravity using; first elution: 6 mL of *n*-hexane to elute PAHs and chlorpyrifos and cypermethrin; second elution: 2 × 3 mL of more polar solvent, dichloromethane. Internal standard (1 mL 20 mg L^{-1}) was added to all fractions prior to nitrogen blow down to a final volume of 1 mL for gas chromatographic analysis.

Accelerated solvent extraction

Extraction cell loading was done in the following sequence: cellulose filter was placed at the bottom of cell, followed by 5 g of silica, another cellulose filter, and finally soil sample (5 g) mixed with diatomaceous earth (Osman et al. 2008). The sample cells were then closed to finger tightness and placed into the carousel of the accelerated solvent extraction system. Two solvents: *n*-hexane and methanol (MeOH) were utilized as extraction solvents. In the first cycle, *n*-hexane was pumped into the cell and was preheated for 2 min to reach the set temperature and pressure followed by a static extraction of 10 min. At the end of the cycle, the pressure was released and the extract was collected in a 60-mL glass vials. The cell was rinsed with fresh solvent (about 80% of extraction cell volume) and purged using pure nitrogen for 1 min. For the second extraction cycle, the sample was extracted again using MeOH under the same

conditions. Extract was collected into a second collection vial. Internal standards (phenantrene, d_{10}, and 5α-cholestane, 1 mL each) were added to the extracts and the volume was reduced to 1 mL prior to gas chromatograph analysis.

Gas chromatographic analysis

In this study, gas chromatographic with flame ionization detector (GC-FID) and gas chromatographic with the electron capture detector (GC-ECD) were used to separate all compounds.

GC-FID analysis

Gas chromatographic separation and identification of PAHs and sterols was performed using an HP6890 series II (Agilent Technologies Inc., Palo Alto, CA, USA) with splitless injection and flame ionization detection. A 30 m × 0.25 mm id × 0.25 μm film thickness HP5-MS capillary column (Agilent technologies) was used to achieve separation of PAHs and sterols with the following temperature program: initial temperature, 50°C; held for 2 min; increased by 18°C min^{-1} to 250°C, increased by 10°C min^{-1} to 310°C; held for 11 min. The detector temperature was set at 310°C. PAH quantification was carried out using a five-point calibration plot containing 5, 10, 25, 50, and 100 mg L^{-1} PAH standard mixtures and 20 mg L^{-1} internal standard (phenantrene, d_{10}). Sterol quantification was carried out using five-point calibration plot containing 10, 25, 50, and 100 mg L^{-1} sterol standard mixtures and 20 mg L^{-1} internal standards (5α-cholestane).

GC-ECD analysis

Separation of chlorpyrifos was achieved using HP7890A gas chromatograph equipped with ^{63}Ni electron capture detectors, GC-ECD (Agilent Technologies Inc., Palo Alto, CA, USA). A 30 m × 0.25 mm id × 0.25 μm film thickness HP5-MS capillary column (Agilent technologies) was used for the quantitative analysis of chlorpyrifos. The injection port and detector temperatures were set at 250°C. The injection volume was 1 μL, and the splitless period following the injection was 2 min. The ECD detector utilized pure N_2 (>99.999%) as a carrier and make-up gas at a controlled constant velocity of 60 mL min^{-1}. The temperature program of the HP5-MS column was set to 150°C for 1 min. then increased by 25°C min^{-1} to 260°C for 8 min. Compounds were identified based on the retention time of the standards and quantified by external standard calibration.

Chemometric approach

Environmental data sets are usually complex and contain a large amount of information with internal relationships among variables, often in a partially hidden structure. The goal of chemometric studies is to display the most significant patterns, looking for possible groupings and sources of data variation, as well as for their temporal and geographical distributions through resolution and modelling of the raw data (Tauler et al. 2004). After data conversion into a single matrix formed by concentration values for each combination of variables and cases, a stepwise statistical approach was used employing the following exploratory techniques: cluster analysis, discriminant analysis, and principal component analysis. Prior to analysis, non-detected values were replaced with half the detection limit (Reimann and Filzmoser 2000). In this study, the XLSTAT2009 software package was employed for multivariate statistical calculations.

Cluster analysis

Cluster analysis, an unsupervised technique, was applied to discover natural groupings within real data in terms of samples similarity. The squared Euclidean distance was always used as the interval measure for clustering using distinct linkage methods: between groups linkage, within groups linkage, and Ward's method. Raw data was computed after standardization based on Z-scores by variable. Cluster analysis groups the objects (cases) into classes (cluster) based on similarities within a class and dissimilarity between different classes. The results of CA help in interpreting the data and indicating patterns (Singh et al. 2004, 2005; Vega et al. 1998).

Discriminant analysis

Discriminant analysis determines the variables that discriminate between two or more naturally occurring groups/clusters. It constructs a discriminant function (DF) for each group (Singh et al. 2004, 2005). DFs were calculated using Eq. 1

$$f(G_i) = k_i + \sum_{j=1}^{n} w_{ij} P_{ij} \quad (1)$$

where i is the number of groups (G), k_i the constant inherent to each group, n the number of parameters used to classify a set of data into a given group, and w_j is the weight coefficient assigned by DF analysis to a given parameter (p_j). In this study, DA was applied to determine whether groups differ with regards to the mean of a variable, and to use that variable to predict group membership. DA was applied to the transform data by using the standard, forward stepwise, and backward stepwise modes. These were used to construct DFs to evaluate variations of the organic contaminants in the river water quality. The identified organic contaminants were the grouping (dependent) variables, while all the measured parameters constitute the independent variables. In the forward stepwise mode, variables were included step-by-step beginning from the most significant variable until no significant changes were obtained. In the backward stepwise mode, variables were removed step-by-step beginning with the less significant variable until no significant changes were obtained.

Principal component analysis

Principal component analysis provides information on the most meaningful parameters that describe the whole data sets rendering data reduction with a minimum loss of original information (Vega et al. 1998). The PCA technique allowed the identification of an association between variables, thus reducing the dimensionality of the data sets. It is a powerful technique for pattern recognition that attempts to explain the variance of a large set of inter-correlated variables and transform them into a smaller set of independent (uncorrelated) variables (principal components). In this study, the Varimax mode of PCA is used, which enables increasing of the weight of the higher factor loading values and reduction of the weight of the lower values. This leads to better understanding of the data structure (Simeonov et al. 2002).

Results and discussion

Descriptive statistic

The data pertaining to the distribution of organic contaminants in the water and soil/sediment samples are tabulated in Table 1 as descriptive statistical parameters. The standard deviation (SD) values related to the distribution of these organic contaminants in soil samples show a very high dispersion around the organic contaminant's concentration. Data on organic contaminant distributions in the two media based on mean concentrations in soil samples show cholesterol as a dominant contaminant with the highest mean concentration of 648.9 μg kg^{-1}, followed by stigmastanol, stigmasterol, and β-sitosterol, at 282.9, 248.2, and 165.6 μg kg^{-1}, respectively. Other with high concentration are benzo[a]anthracene, chlorpyrifos, and cypermethrin at 73.8, 93.3, and 75.7 μg kg^{-1}, respectively. Chlorpyrifos was found at high concentrations (5.06 μg L^{-1}) in water samples, followed by cholesterol, stigmasterol, β-sitosterol, and coprostanol at 1.26, 0.76, 0.57, and 0.47 μg L^{-1}, respectively.

Sites similarity

CA was applied to detect similarity groups between the sampling sites. The data sets were treated (after data scaling by z-transformation) by the Ward's method of linkage with squared Euclidean distance as a measure of similarity. The dendrogram of the locations of different sites along Langat river applied for water data sets are presented in Fig. 1. It shows that the monitoring locations can be grouped into three clusters. Cluster 1 is formed by the sites IL01, IL02, IL07, IL08, IL13, IL18, IL21, and IL24 correspond to the less polluted sites. These stations lie in the rural areas, far from municipal pollution except the station

Table 1 Basic statistical parameter of organic contaminants in water (μg L^{-1}) and soil/sediment (μg kg^{-1}) samples from Langat river basin ($n = 24$)

Organic	Water (μg L^{-1})				Soil/sediment (μg kg^{-1})			
Contaminants	Min	Max	Mean	SD	Min	Max	Mean	SD
Naphthalene	0.025	1.530	0.390	0.544	0.977	80.407	23.571	23.816
Acenaphthalene	0.025	0.344	0.075	0.090	0.977	83.446	17.832	21.017
Acenaphthene	0.025	0.803	0.072	0.172	1.039	45.925	7.056	13.108
Fluorene	0.050	2.483	0.315	0.558	0.072	67.833	10.379	15.758
Pyrene	0.050	1.158	0.269	0.307	1.955	238.893	64.233	74.195
BaA	0.014	1.224	0.227	0.335	0.050	812.786	73.800	178.426
BaP	0.050	0.204	0.062	0.040	2.000	84.972	12.414	23.212
Coprostanol	0.050	1.602	0.465	0.358	2.939	296.308	64.435	68.881
Cholesterol	0.050	5.393	1.255	1.091	27.680	7,233.68	648.908	1,496.61
Stigmasterol	0.050	2.619	0.764	0.604	40.928	996.988	248.233	253.718
B−sitosterol	0.100	5.272	0.569	1.064	4.157	929.378	165.626	239.898
Stigmastanol	0.159	0.485	0.260	0.061	10.129	1,899.49	282.933	489.738
Chlorpyrifos	0.500	14.69	5.057	3.623	3.107	439.020	93.329	107.326
Cypermethrin	0.165	0.165	0.165	0.000	6.450	495.880	75.682	132.802

SD standard deviation, *BaA* benzo[a]anthracene, *BaP* benzo[a]pyrene

IL08 and located in the tributary Batang Labu, near Salak Tinggi, which flows through the areas in fewer pollution activities. The pollution in these areas is mainly from domestic discharges. The cluster 2 formed by the sites IL03, IL04, IL05, IL09, IL10, IL12, 1L14, 1L15, IL16, 1L17, IL20, IL22, and IL25 correspond to the high pollution sites near city area. The results were in an agreement with the study conducted by Juahir (2008) which found that the middle and lower region of this river have experienced urbanization over the past 15 years as the town of Kajang, Cheras, Putrajaya, Teluk Datok, and Teluk Panglima Garang has grown and developed. Stations IL12, IL20, and IL22 are located in a tributary of Batang Benar and Batang Nilai, which carries high pollution flowing through the industrial areas. The cluster 3 formed by sites IL11, IL19, and IL23 correspond to the medium polluted sites except the station IL23 that is located near Nilai city with

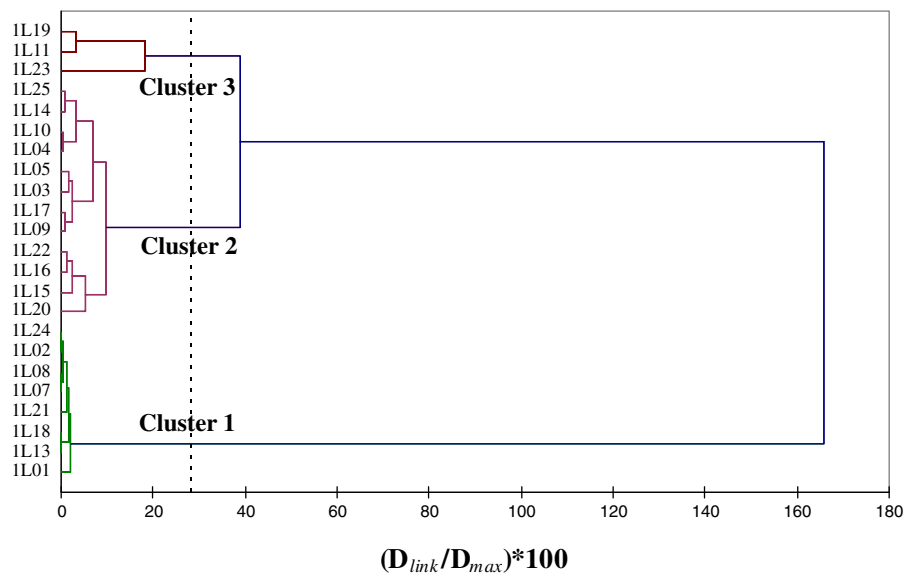

Fig. 1 Dendrogram of clustering of sampling sites according to organic contaminants analyzed in water samples of Langat river and its tributaries using Wards' method

inputs from industries and urban activities. The major polluting industries in Langat river basin are plastic and PVC, engineering products, wood and paper products, and textile and electronic (UPUM 2002).

According to Mokhtar et al. (2002), water pollution has become one of the main problems in the Langat Basin and increased water deficiency is expected to worsen the situation. Water pollution in the urban areas mainly comes from municipal sewerage system and storm-water drainage. Agricultural activities and orchard plantation contributed to a medium polluted area while urban wastewater and industries area contributed to the high amount of the organic pollutants and can be categorize as highly polluted areas in this river basin. The qualitative and quantitative composition of the organic material in the sediment and water column reflects the discharge of anthropogenic contaminants (Schwarzbauer et al. 2000). The CA carry out water samples data indicates that this approach makes possible the design of a future spatial sampling strategy in an optimal way and offers a reliable classification of surface waters in the whole region. For instance, the number of the sampling site could be optimized in such a way that for rapid quality assessment studies only representative sites from each cluster (not all monitoring sites) can be used. Variations of organic contaminants in water and sediment in Langat river were further evaluated using discriminant analysis.

Discriminant analysis for water and sediment samples

Spatial DA was evaluated using standardized data of water and sediment samples. Clearly, the plot of discriminant functions obtained from water and sediment samples (Fig. 2) shows that the pollution comes from three sources: industry and urban socioeconomic, agricultural activity, and urban/domestic sewage. Classification matrices of water data sets obtained from standard, forward stepwise, and backward stepwise modes of DA are shown in Table 2. The standard mode yielded the corresponding correlation matrices assigning 100.0% correctly using 13 variables (Table 2). The forward stepwise mode yielded 66.67% correctly classified using only three discriminant variables (naphthalene, coprostanol, and cholesterol) whereas the percent of cases correctly classified in backward stepwise mode is 87.50% using only four discriminant variables (naphthalene, pyrene, benzo[a]anthracene, coprostanol, and cholesterol). Thus, DA results suggest that naphthalene, pyrene, benzo[a]anthracene, cholesterol, and coprostanol are the significant parameters (Table 3) to discriminate organic contaminants

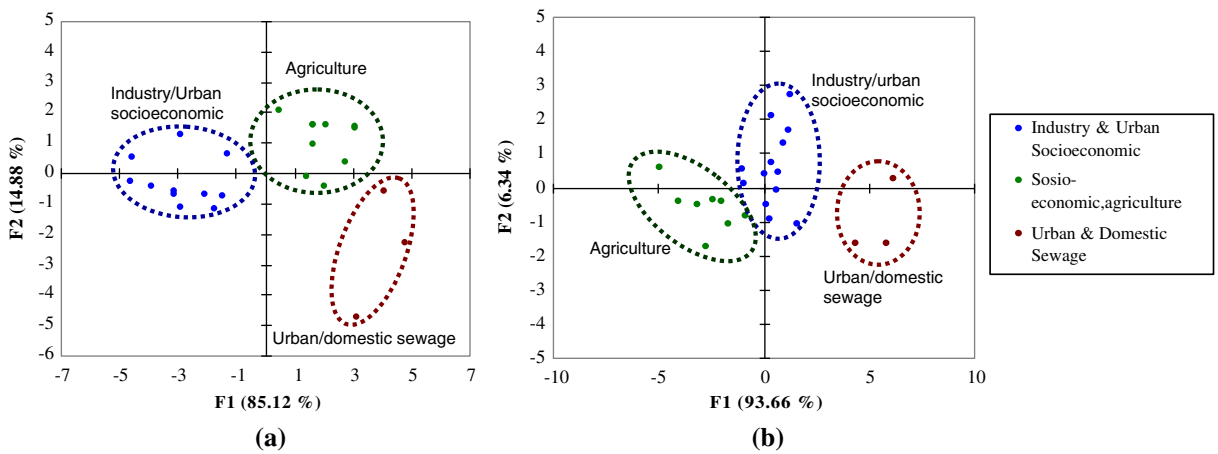

Fig. 2 Plot of discriminant functions showing three categories of pollution sources in Langat river basin based on **a** water samples data sets, **b** soil/sediment samples data sets

Table 2 Classification matrix of DA for all data measurements in water samples ($n = 24$) from Langat river basin

Pollution source	% Correct	Source assigned by DA		
		Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage
Standard DA mode				
Industry and urban socio-economic	100.0	11	0	0
Agricultural and socio-economic	100.0	0	10	0
Urban and domestic sewage	100.0	0	0	3
Total	100.0	11	10	3
Forward stepwise DA mode				
Industry and urban socio-economic	54.55	6	4	1
Agricultural and socio-economic	90.00	1	9	0
Urban and domestic sewage	33.33	0	2	1
Total	66.67	7	15	2
Backward stepwise DA mode				
Industry and urban socio-economic	90.91	10	1	0
Agricultural and socio-economic	100.0	0	10	0
Urban and domestic sewage	33.33	0	2	1
Total	87.50	10	13	1

detected in water samples between the sampling stations.

Same procedures were applied to the data sets of soil/sediment samples from this river and the results were tabulated in Tables 4 and 5. The standard mode yielded the correlation matrices assigning 95.8% correctly using 14 variables (Table 4). The forward stepwise mode yielded 58.33% correctly using only one discriminant variable (acenaphthalene) and the percent of cases correctly classified in backward stepwise mode is 95.8% using nine discriminant variables (naphthalene, acenaphthene, fluorene, pyrene, anthracene, benzo[a]pyrene, cholesterol, stigmasterol, and chlorpyrifos). Higher concentration of naphthalene was obtained in water obtained from the sampling stations which the pollution source mainly comes from industrial and urban activities. Same pattern was observed for pyrene and benzo[a]anthracene. However, naphthalene was not selected as a discriminant variable for sediment samples. This phenomenon may be due to the high solubility of this compound in water, so it remained in water and not accumulated in sediment. On the other hand, the high molecular weight PAHs (benzo[a]pyrene and pyrene) tend to accumulate in sediment as these compounds were found higher at sampling stations located near urban and industrial area. Zhang et al. (2006) reported the levels of PAHs in soils in urban areas was approximately two to ten times higher than those in rural areas. The concentration of coprostanol and cholesterol were higher in sediment samples representing sewage pollution from domestic and urban area. Several studies (Carreira et al. 2004; Chan et al. 1998; Isobe et al. 2002) demonstrated fecal sterols such as coprostanol and cholesterol to be useful sewage markers for sewage pollution. The concentration of chlorpyrifos and cypermethrin were significantly higher at sampling sites near agricultural area compared to other areas. Study conducted by Goncalves et al. (2006) found that the occurrence of pesticides (including chlorpyrifos) in the soil sample from horticulture area was successfully interpreted by chemometric techniques.

Principal component analysis

The principal component analysis was applied on the data of organic contaminants for factor loading in each medium (water and sediment) to identify the spatial sources of pollution in Langat river basin. According to the eigenvalue criterion, only the PCs with eigenvalue greater than one are considered important. This criterion is based on the fact that the average eigenvalue of the autoscaled data is just one (Kowalkowski et al. 2006). For the water data sets, with the eigenvalue criteria (eigenvalue >1), the factor analysis with varimax

Table 3 Classification function coefficients of DA for all data measurements in water samples ($n = 24$) from Langat river basin

Parameters	Standard mode			Forward stepwise mode			Backward stepwise mode		
	Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage	Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage	Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage
Naphthalene	5.687	−4.552	−3.416	4.747	0.565	0.698	11.112	1.816	1.678
Acenaphthalene	55.946	−24.334	−92.167						
Acenaphthene	67.396	59.166	37.135						
Fluorene	20.984	15.361	9.183						
Pyrene	25.411	−5.267	−8.547				23.548	4.904	6.067
BaA	−60.952	−3.045	29.249				−25.033	−4.910	−3.759
BaP	207.722	348.909	389.907						
Coprostanol	24.316	38.934	61.914	4.872	1.329	13.176	1.943	0.744	12.650
Cholesterol	9.039	−2.549	−13.888	0.756	0.464	−2.767	5.572	1.400	−2.126
Stigmasterol	−34.910	−43.248	−39.242						
β−sitosterol	−4.138	−3.089	−3.470						
Stigmastanol	438.890	514.773	517.784						
Chlorpyrifos	−1.279	−2.353	−3.326						
Constant	−66.575	−67.421	−74.219	−4.825	−1.240	−6.078	−11.618	−1.523	−6.607

BaA benzo[a]anthracene, *BaP* benzo[a]pyrene

Table 4 Classification matrix of DA for all data measurements in soil/sediment samples ($n = 24$) from Langat river basin

Pollution source	% Correct	Source assigned by DA		
		Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage
Standard DA mode				
Industry and urban socio-economic	100.0	13	0	0
Agricultural and socio-economic	87.5	1	7	0
Urban and domestic sewage	100.0	0	0	3
Total	95.8	14	7	3
Forward stepwise DA mode				
Industry and urban socio-economic	100.0	13	0	0
Agricultural and socio-economic	0.00	8	0	0
Urban and domestic sewage	33.33	2	0	1
Total	58.33	23	0	1
Backward stepwise DA mode				
Industry and urban socio-economic	100.0	13	0	0
Agricultural and socio-economic	87.5	1	7	0
Urban and domestic sewage	100.0	0	0	3
Total	95.8	14	7	3

rotation resulted in four varifactors (VFs) comprised of 79.04% total variance. Table 6 shows that among four VFs, VF1 accounts for 35.3% of the total variance showing high positive loading on coprostanol and cholesterol. High positive loading on coprostanol is suspected to originate from sewage contamination because this compound is one of the sterols found in human feces and was originally proposed as a potential marker of human sewage contamination (Carreira et al. 2004; Chan et al. 1998). Coprostanol has been widely used as a marker of fecal pollution (Mudge and Seguel 1999; Takada and Eganhouse 1998) because it is produced in the digestive tracts of humans and higher vertebrates by the microbial reduction of cholesterol. It comprises 40–60% of the total fecal sterols excreted in human wastes (Brown et al. 1980; Gilli et al. 2006; Leeming and Nichols 1996).

VF2 contributed 21.0% of the total variance shows high positive loading on stigmasterol, β-sitosterol, and stigmastanol. According to Leeming and Nichols (1996) and Saim et al. (2009), the presence of this sterols is much related to the livestock farming activities. The same study found that the source-specificity of fecal sterols is caused by a combination of three main factors; animal's diet, biosynthesized by higher animals and discharged to the digestive tract and anaerobic bacteria in the digestive tract of some animal biohydrogenate sterols to stanols of various isomeric configurations. The combination of these three factors determines "the sterol fingerprint" of any animals. The results concurrent to the DOE (2006) whereby sewage treatment plants recorded the highest number of pollution in Langat river basin, followed by the manufacturing sector, livestock farming, and agro-based industries (DOE 2006). VF3 and VF4 account for 12.9% and 9.8% of the total variance, respectively, was predominated by PAHs compounds (acenaphthene, pyrene, benzo[a]anthracene, naphthalene, fluorene, and benzo[a]pyrene). These two factors indicate their origin from vehicle emission and combustion sources (Li et al. 2006; Omar et al. 2002; Tahir et al. 2006). Naphthalene is ubiquitous environmental pollutant and high level of naphthalene occurs at certain industrial workplaces (Preuss et al. 2003).

Table 7 brings out the principal component loadings for soil/sediment sample having a total variance of 77.6%, where five factors were extracted with contributions of 26.4%, 12.8%, 12.4%, 13.8%, and 12.25%, respectively. PCA using varimax rotation was applied for factor loading in this sample. The first varimax factor, VF1 with the highest total variance (26.4%) showed the high loadings for pyrene, stigmasterol, β-sitosterol, and stigmastanol. These organic contaminants

Table 5 Classification function coefficients of DA for all data measurements from Langat river basin soil/sediment samples ($n = 24$)

Parameters	Standard mode			Forward stepwise mode			Backward stepwise mode		
	Industry and urban socio-economic	Agricultural and socio-economic	Urban and domestic sewage	Industry and urban socio-economic	Agricultural and socio-economic sewage	Urban and domestic	Industry and urban socio-economic sewage	Agricultural and socio-economic	Urban and domestic
Naphthalene	0.135	0.073	0.125						
Acenaphthalene	0.021	−0.060	0.239	0.042	0.034	0.128	0.065	−0.001	0.262
Acenaphthene	0.495	−0.112	0.937				0.224	−0.165	0.796
Fluorene	0.205	−0.173	0.774				0.135	−0.117	0.644
Pyrene	0.170	−0.150	0.441				0.071	−0.089	0.297
Bezo[a]anthracene	−0.048	0.042	−0.119				−0.020	0.026	−0.088
Benzo[a]pyrene	0.145	−0.344	0.598				0.121	−0.122	0.416
Coprostanol	−0.009	0.004	0.054						
Cholesterol	0.002	−0.004	0.012				0.002	−0.001	0.008
Stigmasterol	−0.006	0.018	−0.091				−0.007	0.011	−0.051
β−sitosterol	−0.005	−0.024	0.032						
Stigmastanol	−0.006	0.010	−0.012						
Chlorpyrifos	−0.072	0.120	−0.266				−0.041	0.064	−0.186
Cypermethrin	−0.004	0.025	−0.018						
Constant	−6.188	−5.866	−32.516	−0.927	−1.304	−4.992	−3.373	−3.274	−25.486

Table 6 Loadings of organic contaminants after Varimax rotation (PC extracted four factors) for water samples collected from 24 sampling stations of Langat river basin (high loadings >0.75 are shown in bold; moderate loading 0.5–0.75 in italic bold)

Parameters	VF1	VF2	VF3	VF4
Naphthalene	0.221	−0.114	0.194	**0.846**
Acenaphthalene	*0.610*	−0.083	0.443	0.375
Acenaphthene	−0.165	−0.139	**0.760**	−0.178
Fluorene	0.055	−0.038	0.082	**0.774**
Pyrene	0.234	0.300	**0.761**	0.243
BaA	*0.574*	−0.041	**0.749**	0.113
BaP	0.125	0.077	−0.130	***0.722***
Coprostanol	**0.892**	0.221	0.084	0.075
Cholesterol	**0.932**	0.165	0.090	0.103
Stigmasterol	*0.512*	**0.787**	0.023	0.115
β−sitosterol	−0.006	**0.952**	−0.043	0.013
Stigmastanol	0.147	**0.829**	0.035	−0.400
Chlorpyrifos	0.109	*0.669*	0.555	0.139
Eigenvalue	4.595	2.735	1.675	1.269
Explained variance (%)	35.349	21.038	12.888	9.761
Cumulative (%)	35.349	56.387	69.275	79.036

BaA benzo[a]anthracene, *BaP* benzo[a]pyrene

originate from anthropogenic activities related to vehicle emission (Li et al. 2006) and livestock farming waste (Leeming and Nichols 1996). Meanwhile, moderate loading of chlorpyrifos was also observed (0.649). This compound is correlated to the agricultural activities especially oil palm plantation within the Langat river basin (Juahir et al. 2009). Varifactor 2 (VF2), with a total of 12.8% variance has the highest loadings for acenaphthene and benzo[a]anthracene, manifesting the common wood burning source (Khalili et al. 1995; Rogge et al. 1998). Varifactor 3 showed strong loading of coprostanol and acenaphthalene which suggested as chemical markers for sewage contamination and vehicle markers, respectively (Isobe et al. 2004; Li et al. 2006). The last two varifactors (VF4 and VF5), with a total variance of 13.8% and 12.2% have high loading of cholesterol and benzo[a]pyrene, respectively. These factors are conceived to originate from domestic or urban sewage and incomplete burning activities. The Department of Environment, Malaysia (DOE 2006), has reported that water pollution in the urban areas is mainly related to the municipal sewerage system and storm-water drainage. High concentration of PAHs was observed in water samples (benzo[a]anthracene, naphthalene, and

Table 7 Loadings of organic contaminants after Varimax rotation (PC extracted five factors) for soil/sediment samples collected from 24 sampling stations of Langat river basin (strong loadings >0.75 are shown in bold; moderate loading 0.5–0.75 in italic bold)

Parameters	VF1	VF2	VF3	VF4	VF5
Naphthalene	−0.123	−0.174	−0.038	***0.697***	0.447
Acenaphthalene	−0.073	0.062	**0.838**	−0.155	0.095
Acenaphthene	−0.052	**0.856**	0.252	−0.084	−0.150
Fluorene	***0.698***	−0.316	0.225	−0.026	−0.164
Pyrene	**0.774**	0.185	0.016	−0.013	0.416
BaA	0.077	**0.809**	−0.151	0.221	0.278
BaP	0.056	0.102	0.027	0.088	**0.911**
Coprostanol	0.338	0.064	**0.829**	0.166	−0.038
Cholesterol	0.235	0.136	0.002	**0.887**	−0.080
Stigmasterol	**0.734**	0.156	0.088	0.575	0.064
β−sitosterol	**0.911**	0.079	−0.028	0.083	−0.083
Stigmastanol	**0.748**	−0.106	0.189	0.380	0.075
Chlorpyrifos	***0.649***	−0.296	0.041	−0.155	0.534
Cypermethrin	0.241	0.281	−0.398	−0.195	0.249
Eigenvalue	4.295	1.977	1.753	1.524	1.322
Explained variance (%)	26.411	12.832	12.376	13.808	12.219
Cumulative (%)	26.411	39.243	51.619	65.427	77.647

BaA benzo[a]anthracene, *BaP* benzo[a]pyrene

pyrene) and sediment samples (benzo[a]pyrene and pyrene) collected at sampling sites near industrial area.

Conclusion

In this study, chemometric techniques (cluster analysis, discriminant analysis, and principal component analysis) were successfully applied to evaluate the variations of organic contaminants (polycyclic aromatic hydrocarbons, chlorpyrifos, cypermethrin, and sterols) within Langat river basin. The results suggest that the profile of organic contaminants differs depending on the sources of pollution. This could lead to the development of chemometric techniques to find similarities and differences in the sources of the data. Hierarchical cluster analysis group sampling sites into three clusters of similar organic contaminant characteristics. Discriminant's analysis yielded an important data reduction, as it used only ten parameters (naphthalene, acenaphthene, benzo[a]anthracene, benzo[a]pyrene, coprostanol, cholesterol, stigmasterol, β-sitosterol, stigmastanol, and chlorpyrifos) affording more than 90% correct assignations. Principal component analysis is useful in extracting and identifying the factors responsible for the variations in organic contaminants at different sampling sites. Results from PCA indicated that sterols (coprostanol, cholesterol, stigmasterol, β-sitosterol, and stigmastanol) are strongly correlated to domestic and urban sewage while PAHs (naphthalene, acenaphthene, pyrene, benzo[a]anthracene, and benzo[a]pyrene) from industrial and urban activities and chlorpyrifos correlated to samples nearby agricultural sites. These techniques provided a more objective interpretation of the results and enhanced the use of selected organic contaminants as source tracers for organic contamination environmental compartments (water and sediment).

Acknowledgements The authors would like to acknowledge a financial support obtained from Ministry of Higher Education (MOHE), Malaysia, for this project (project number: 600-IRDC/ST/FRGS 5/3/1318).

References

Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software, 20*, 1263–1271.

Astel, A., Tsakovski, S., Barbieri, P., & Simeonov, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research, 41*(19), 4566–4578.

Boruvka, L., Vecek, O., & Jenlika (2005). Principal component analysis as a tool to indicate the origin of potentially toxic elements in soil. *Geoderma, 128*, 289–300.

Brodnjak-Voncina, D., Dobcnik, D., Novic, M., & Zupan, J. (2002). Chemometrics characterisation of the quality of river water. *Analytica Chimica Acta, 462,* 87–100.

Brown, S. D., Skogerboe, R. K., & Kowalski, B. R. (1980). Pattern recognition assessment of water quality data: Coal strip mine drainage. *Chemosphere, 9*, 265–276.

Carreira, R. S., Wagener, A. L. R., & Readman, J. W. (2004). Sterols as markers of sewage contamination in a tropical urban estuary (Guanabara Bay, Brazil): Space–time variations. *Estuarine, Coastal and Shelf Science, 60*, 587–598.

Chan, K.-H., Lam, M. H. W., Yeung, H.-Y., & Chiu, T. K. T. (1998). Application of sedimentary fecal stanol and sterols in tracing sewage pollution in costal waters. *Water Research, 32*(1), 225–235.

Department of Environmental Malaysia, DOE (2006). Malaysia environmental quality report 2006 in Impak, Kuala Lumpur: Ministry of Science. *Technology and Environment, 3*, 1–16.

Gilli, G., Rovere, R., Traversi, D., Schiliro, T., & Pignata, C. (2006). Faecal sterols determination in wastewater and surface water. *Journal of Chromatography B, 843*, 120–124.

Goncalves, C., Carvalho, J. J., Azenha, M. A., & Alpendurada, M. F. (2006). Optimization of supercritical fluid extraction of pesticide residues in soil by means of central composite design and analysis by gas chromatography–tandem mass spectrometry. *Journal of Chromatography A, 1110*, 6–14.

Isobe, K. O., Tarao, M., Zakaria, M., Chiem, N. H., Minhle, Y., & Takada, H. (2002). Quantitative application of fecal sterols using gas-chromatography–mass spectrometry to investigate fecal pollution in tropical water: Western Malaysia and Mekong Delta, Vietnam. *Environmental Science & Technology, 36*, 4497–4507.

Isobe, K. O., Tarao, M., Chiem, N. H., & Minh, L. Y. (2004). Effect of environmental factors on the relationship between concentrations of coprostanol and fecal indicator bacteria in tropical (Mekong Delta) and temperate (Tokyo) freshwaters. *Applied and Environmental Microbiology, 70*(2), 814–821.

Juahir, H. (2008). *Water Quality Data Analysis and Modeling at Langat River Basin*. PhD dissertation, Universiti Putra Malaysia.

Juahir, H., Zain, S. M., Khan, R. A., Yusoff, M. K., Mokhtar, M. B., & Toriman, M. E. (2009). Using chemometrics in assessing Langat River water quality and designing a cost-effective water sampling strategy. *Maejo International Journal of Science and Technology, 3*(01), 26–42.

Khalili, N. R., Scheff, P. A., & Holsen, T. M. (1995). PAHs source finger prints for coke ovens, diesel and gasoline engines, highway tunnels and wood combustion emm. *Atmospheric Environment, 29*, 533–542.

Kowalkowski, T., Zbytniewski, R., Szpejna, J., & Buszewski, B. (2006). Application of chemometrics in river water classification. *Water Research, 40*, 744–752.

Leeming, R., & Nichols, P. (1996). Concentrations of coprostanol that correspond to existing bacterial indicator guideline limits. *Water Research, 30*(12), 2997–3006.

Li, J., Zhang, G., Li, X. D., Qi, S. H., Liu, G. Q., & Peng, X. Z. (2006). Source seasonality of polycyclic aromatic hydrocarbons (PAHs) in a subtropical city, Guangzhou, South China. *Science of the Total Environment, 355*, 145–155.

Mohamed, A. F., Wan Yaacob, W. Z., Taha, M. R., & Shamsudin, A. R. (2009). Groundwater and soil vulnerability in the Langat Basin Malaysia. *European Journal of Scientific Research, 27*(4), 628–635.

Mokhtar, M., Mohamed, A. F., & Idrus, S. (2002). Preliminary survey of groundwater resources in the Langat basin: Towards a sustainable healthy ecosystem. In M. Mokhtar, S. Idrus, A. F. Mohamed, A. H. A. Shah, S. Aziz & A. G. Aziz (Eds.), *Proceedings Langat basin research symposium 2001*. Institute for Environmental and Development, Universiti Kebangsaan Malaysia, Bangi.

Mudge, S. M., & Seguel, C. G. (1999). Organic contamination of San Vicente Bay, Chile. *Marine Pollution Bulletin, 38*(11), 1011–1021.

Omar, N. Y. M. J., Abas, M. R. B., Ketuly, K. A., & Tahir, N. M. (2002). Concentrations of PAHs in atmospheric particles (PM-10) and roadside soil particles collected in Kuala Lumpur, Malaysia. *Atmospheric Environment, 36*, 247–254.

Osman, R., Saim, N., & Abdullah, M. P. (2008). Selective accelerated solvent extraction for the analysis of polycyclic aromatic hydrocarbons and sterols from soil. *The Malaysian Journal of Analytical Sciences, 12*(2), 352–356.

Osman, R., Saim, N., & Abdullah, M. P. (2009). Simultaneous extraction of organic compounds with a wide polarity range in water using solid phase extraction technique. *Research Journal of Chemistry and Environment, 13*(3), 7–18.

Preuss, R., Angerer, J., & Drexler, H. (2003). Naphthalene—An environmental and occupational toxicant. *International Archives of Occupational and Environmental Health, 76*, 556–576.

Reimann, C., & Filzmoser, P. (2000). Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology, 39*, 1001–1014.

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., & Simoneit, B. R. T. (1998). Sources of fine organic aerosol. 9. Pine, oak, and synthetic log combustion in residential fireplaces. *Environmental Science and Technology, 32*, 13–22.

Saim, N., Osman, R., Spian, D. R. S. A., Jaafar, M. Z., Juahir, H., Abdullah, M. P., et al. (2009). Chemometric approach to validating faecal sterols as source tracer for faecal contamination in water. *Water Research, 43*, 5023–5030.

Schwarzbauer, J., Littke, R., & Weigelt, V. (2000). Identification of specific organic contaminants for estimating the contribution of the Elbe river to the pollution of the German Bight. *Organic Geochemistry, 31*, 1713–1731.

Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River basin Japan. *Environmental Modelling and Software, 22*(4), 464–475.

Simeonov, V., Einax, J. W., Stanimirova, I., & Kraft, J. (2002). Environmetric modeling and interpretation of river water monitoring data. *Analytical and Bioanalytical Chemistry, 374*(5), 898–905.

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., et al. (2003). Assessment of the surface water quality in Northern Greece. *Water Research, 37*, 4119–4124.

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. *Water Research, 38*, 3980–3992.

Singh, K. P., Malik, A., Singh, V. K., Mohan, D., & Sinha, S. (2005). Chemometric analysis of groundwater quality data of alluvial aquifer of Gangetic plain, North India. *Analytica Chimica Acta, 550*, 82–91.

Tahir, N. M., Seng, T. H., Ariffin, M., Suratman, S., & Hoe, L. S. (2006). Concentration and distribution of PAHs in soils affected by grassland fire. *The Malaysian Journal of Analytical Sciences, 10*(1), 41–46.

Takada, H., & Eganhouse, R. P. (1998). Molecular markers of anthropogenic waste. In R. A. Meyer (Ed.), *Encyclopedia of Environmental Analysis and Remediation* (pp. 2883–2940). New York: Wiley.

Tauler, R., Peré-Trepat, E., Lacorte, S., & Barceló, D. (2004). *Chemometrics Modelling of Environmental Data*. Department of Environmental Chemistry, Institute of Chemical and Environmental Research IIQAB-CSIC, Spain.

University Malaya Consultancy Unit (UPUM) (2002). *Program Pencegahan Pencemaran dan Peningkatan Kualiti Air Sungai Lang at; Final Draft Report*.

Vega, M., Pardom, R., Barrado, E., & Ddebaan, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research, 32*(12), 3581–3592.

Zhang, H. B., Wong, Y. M. L. H., Zhao, Q. G., & Zhang, G. L. (2006). Distributions and concentrations of PAHs in Hong Kong soils. *Environmental Pollution, 141*, 107–114.