

## Characterization of Hopanes in Surface Sediments from Southwest corner of South China Sea: Application of PCA Utilizing of Individual Hopanes (Pencirian hopanes di dalam mendakan permukaan di sudut Barat Daya Laut China Selatan: Aplikasi penggunaan PCA daripada individu hopane)

<sup>1</sup>Ismail N., <sup>1</sup>Zakaria M.P., <sup>2</sup>Abdullah A.H., <sup>3</sup>Mohamed C.A.R., <sup>1</sup>Sukri N.S., <sup>1</sup>Manan N., <sup>1</sup>Masood N.A., <sup>4</sup>Juahir H.

<sup>1</sup>Center of Excellence for Environmental Forensics, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

<sup>2</sup>Department of Chemistry, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

<sup>3</sup>Marine Ecosystem Research Centre, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

<sup>4</sup>Department of Environmental Science, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

Received 6 July 2013; accepted 5 September 2013

### ABSTRACT

Oil pollution has been introduced into the aquatic environment of the South China Sea through anthropogenic activities such as discharging of oil during extraction, transportation and consumption. In order to determine the source of oil pollution in South China Sea, 30 surface sediment samples were collected in 2008 and were analyzed for pentacyclic triterpanes (hopane) by gas chromatography mass spectrometer (GCMS) with  $m/z$  191. This study investigates the utility and limitation of individual hopane in tracing the sources of oil pollution by using Principle Component Analysis (PCA). Characterization the utility of hopane in determining the source of oil pollution has been done by integrating diagnostic ratios of  $C_{29}/C_{30}$  and  $\sum C_{31}-C_{35}/C_{30}$  with PCA. PCA has classified all target components into 3 major principle components prior to determine their importance in identifying the source of oil pollution. From the rotated component pattern, first principle component (63.24% of variability) loaded by 9 target compounds which namely Ts, Tm, 17 $\alpha$ (H), 21 $\beta$ (H)- norhopane, 17 $\alpha$ (H), 21 $\beta$ (H)- hopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- bishomohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- trishomohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- tetrakishomohopane and 22S, 17 $\alpha$ (H) and 21 $\beta$ (H)- pentakishopane. Those individual hopanes listed in PC1 were excellent in tracing both oil pollution as all of them degrade less and generally unaffected by even severe biodegradation. This findings give better understanding for non consideration of both 17 $\beta$ (H), 21 $\alpha$ (H)- norhopane, 17 $\beta$ (H), 21 $\alpha$ (H)- hopane in determining the origin of oil spill.

**Keywords:** Pentacyclic triterpane, oil pollution sources, surface sediment, South China Sea, Principle Component Analysis

### ABSTRAK

Pencemaran minyak telah diperkenalkan ke perairan Laut Cina Selatan melalui aktiviti antropogenik seperti pelepasan minyak semasa pengekstrakan, pengangkutan dan penggunaan. Bagi menentukan punca pencemaran minyak dalam Laut Cina selatan, 30 sampel permukaan mendakan telah dikutip pada tahun 2008 dan dianalisis untuk pentacyclic triterpane (hopane) dengan menggunakan kromatografi gas spectrometer jisim (GCMS) dengan  $m/z$  191. Pelajaran ini menyiasat kepentingan dan had bagi hopane individu dalam mengesan punca pencemaran minyak dengan menggunakan analisis prinsip komponen (PCA). Sebelum analisis lanjutan diteruskan, 4 sampel mendakan permukaan telah di kecualikan kerana nilai ekstremnya. Keempat-empat outlier (SF02, SF06, SF08 dan SF22) terletak melebihi kuartil ketiga analisis rawatan outlier. Pencirian kepentingan hopane dalam menentukan punca pencemaran minyak telah dilakukan dengan mengintegrasikan

*nisbah diagnostic  $C_{29}/C_{30}$  and  $\Sigma C_{31}-C_{35}/C_{30}$  dengan PCA. PCA telah mengelaskan kesemua komponen target kepada 3 komponen prinsip utama dalam menentukan kepentingan mereka dalam menentukan punca pencemaran minyak. Daripada corak komponen terputar, komponen prinsip pertama (63.24% kepelbagaian) terbeban oleh 9 kompon target seperti Ts, Tm, 17 $\alpha$ (H), 21 $\beta$ (H)- norhopane, 17 $\alpha$ (H), 21 $\beta$ (H)- hopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- bishomohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- trishomohopane, 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- tetrakishomohopane and 22S, 17 $\alpha$ (H) and 21 $\beta$ (H)- pentakishopane. Kesemua hopane individu yang tersenarai dalam PC1 adalah cemerlang dalam mengesan pencemaran minyak kerana kesemua mereka kurang merendah dan secara umumnya tidak terpengaruh oleh biodegradasi yang teruk. Penemuan ini memberikan kefahaman untuk tidak mengambilkira kedua-dua dalam menentukan asal tumpahan minyak.*

**Kata kunci:** Pentacyclic triterpane, punca pencemaran minyak, mendakan permukaan, Laut Cina Selatan, Analisis Prinsip Komponen

## INTRODUCTION

A rapid economic development of many countries since World War II has caused a considerable increase in marine transportation of raw material, especially of crude oil and in offshore activities. However, a significant amount of oil comes into the sea from operational discharges of ships as well as from incidents such as collisions, groundings and contacts. Offshore exploration and exploitation of oil and gas is connected with the danger of blow-out and major spills. In other hand, economic development in many parts of the world such as in the South East Asia region will exert tremendous pressures on the marine pollution especially organic pollutant from land- and ship-based sources. The bulk of world trade, in tonnage terms, is transported by ships and will remain so for many years. This is particularly important in highly congested and confined sea lanes with high biodiversity such as in South China Sea. South China Sea has an area of approximately 3.3 million km<sup>2</sup> with 270 million people living along its shores. The most important countries surrounding the South China Sea is China in the North, The Philippines in the East, Indonesia to the South area, Malaysia in the South-west and Vietnam in the West (Morton and Blackmore, 2001). The South China Sea has been threatened by a high population growth rate, urbanization, eco-tourism, habitat modification and significantly petroleum pollution. In a report published in 2002 by National Research Council of the U.S. National Academy of Sciences, discharges during petroleum extraction or production tend to be restricted to areas of exploration and extraction and are mostly due to the release of "produced waters" (water extracted with petroleum from the reservoir). This is a common phenomenon in areas where there is an extensive oil exploration such as in the South China Sea (Chandru et al., 2008 and Asuquo, 1991). These contribute about 5% of the petroleum reaching the sea from human sources whereas spillage of petroleum during transportation, refinement, and distribution are most common along shipping routes and pipelines and make up about 22% of human-caused petroleum inputs. Discharges during petroleum consumption (i.e., use of automobiles and boats) tend to be small in quantity but are so numerous and widespread that they contribute the vast majority (about 70%) of human-caused petroleum pollution in the sea.

Biological markers or molecular markers are one of the most important hydrocarbons group in petroleum used for chemical fingerprinting. They are complex molecules derived from formerly living organism (Wang et al., 2007). Biomarkers found in crude oils, rocks and sediments have little or no changes in structure from their parent

biochemicals, so-called biogenic precursors (e.g., steroids and terpanoids), and found living organisms (Wang et al., 2007; Peters and Moldowan, 1993).

Excellent reviews on the fundamentals of biomarker characterization, their application in petroleum geochemistry, and interpretation of biomarker data for oil exploration and production were published in 1993 by Peters and Moldowan. Fully updated and expanded editions provide a comprehensive account of the role that biomarker technologies play both in petroleum exploration and in understanding earth history and processes, including environmental applications (Peters et al., 2005). More recently, Wang et al. (2006) and Chandru et al. (2008) have reviewed the environmental applications of biomarker fingerprinting. Detail discussion on the biomarker concept, criteria and information has been extensively reviewed in Wang et al. (2007). Biomarker that is used in this study is hopane.

Hopanes are pentacyclic triterpanes commonly containing 27-35 carbon atoms in a naphthenic structure composed of four six-membered rings and one five-membered ring (Wang et al., 2007). Hopanes with the  $17\alpha(\text{H}),21\beta(\text{H})$  configuration in the range C27 – C35 are characteristic of petroleum because of their large abundance and thermodynamic stability compared to other epimeric ( $\beta\beta$ ,  $\alpha\alpha$ ) series. Previously, hopanes were considered to exist as three stereoisomers:  $17\alpha(\text{H}),21\beta(\text{H})$ - Hopane,  $17\beta(\text{H}),21\beta(\text{H})$ -Hopane and  $17\beta(\text{H}),21\alpha(\text{H})$ -Hopane (Peters and Moldowan, 1993; Waples and Machihara 1991).

## MATERIALS AND METHODS

### Sampling collection

Figure 1 and Table 1 showed the locations and sampling details in this study. Thirty core sediments were collected from the location in the east coast of Peninsular Malaysia area onward to South China Sea region in June 2008. For the purpose of this study we only focused onto the surface, intermediate and ground core sediment layers which were taken by using pre-cleaned stainless steel scoop and placed on a stainless steel pan. The sediment samples were then transported on an ice to the laboratory and stored at  $-20^{\circ}\text{C}$  until further analysis.

### Chemical analysis

Organic solvents such as methanol, acetone, isooctane, n-hexane, dichloromethane (DCM) and authentic standard solutions for hopanes were purchased from Wako Pure Chemical (Japan), Chiron (Norway) and Sigma (US). DCM, n-hexanes and isooctane were distilled before use. Glassware was washed by using a series of soap, tap water and ultra pure water. Then, it was rinsed respectively with methanol, acetone and n-hexane (HPLC Grade) and baked  $60^{\circ}\text{C}$  for several hours in order to remove any organic contaminants (Chandru et al., 2007).

The hopanes standard mixtures contains  $17\beta(\text{H}),21\alpha(\text{H})$ - norhopane (C29 17b),  $17\alpha(\text{H})$ -22,29,30-trisnorhopane (Tm),  $17\alpha(\text{H}),21\beta(\text{H})$ -hopane (C30 17 $\alpha$ ),  $17\beta(\text{H}),21\alpha(\text{H})$ -hopane (C30 17b) and for IIS  $17\beta,21(\text{H})\beta$ -hopane was used. Activated silica gel was prepared by baking at  $400^{\circ}\text{C}$  for 4 hours, cooled and deactivated with 5% (w/w) distilled  $\text{H}_2\text{O}$ .

Twenty grams of sediment samples were accurately measured, dried with  $\text{Na}_2\text{SO}_4$  and extracted by using Soxhlet extraction. The extract was purified and fractionated by method described elsewhere in Zakaria et al. (2000). The solution was transferred onto the top of a 5%  $\text{H}_2\text{O}$  deactivated silica gel column chromatography. Hydrocarbons ranged from alkanes to PAHs, with seven rings were eluted with 20 ml of DCM/n-hexane (1:3, v/v). The hydrocarbon was fractionated further with fully activated silica gel column chromatography to obtain hopane fractions by eluting 4 ml n-hexane. Hopanes and was detected using mass spectrometer integrated with gas chromatograph (GC–MS). Details of quantitative and qualitative measures for GC–MS conditions were explained (Zakaria et al., 2000, 2001 *a,b*).

### Statistical analysis

Before achieving the objectives of this study, statistical analysis needs to be done in order to treat and discover its underlying causes, patterns, relationships, and trends.

In order to quantitatively analyze and significantly confirm the relationships among hopane and total hopanes concentrations, the Pearson correlation analysis has been applied to the correlation value. The Pearson correlation coefficient is a typical parameter used to identify the strength of linear relation between the pairs of parameters by calculating a summary index.

In an experiment that involved a complex dataset, the expression of thousands of compound is measured across many conditions such as individual compound in that particular composition. Therefore, it becomes impossible to make a visual inspection of the relationship between individual compounds in such a complex matrix. The only way to make sense of this data is by decreasing its dimensionality. Several data decomposition techniques are applicable for this purpose for example Principal Components Analysis (PCA) which is among those redundant techniques that reduces the data into two dimensions by applying a covariance analysis between factors where it provides information on the most meaningful parameters that represent the whole data with minimum percentage of its original data (Juahir et al., 2009). As such, it is suitable for data sets in multiple dimensions, such as a large experiment in sediment layers.

In this research, descriptive statistic has been applied on the data set as a purpose to check in detail for every single extreme value of generated data. This extreme value is also best known as outlier where it would be an observation which deviates so much from other observations as to arouse suspicious that was generated by a different mechanism (Hawkins., 1980). An inspection of a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between the outliers and the group of inliers, as measured on some suitably standardized scale. As an impact, outlier would give different value on mean and standard deviation as compared to the value it supposed to have.

There are several other useful measures where percentiles are points that divide the data. The 25<sup>th</sup> percentile of first quartile, denoted as  $Q_1$ , marks the point below which 25% of the data lie. The 75<sup>th</sup> percentile or third quartile denoted as  $Q_3$  where it is defined similarly (Buncher and Tsay, 2006). The method of identifying the outliers is due to Mosteller and Tukey (1977). This method has been performed by identifying outside values as those that are less than  $Q_1 - (1.5 \times \text{IQR})$  or greater than  $Q_3 + (1.5 \times \text{IQR})$ , where  $Q_1$  and  $Q_3$  are the lower and upper quartiles and  $\text{IQR} = Q_3 - Q_1$ , the inter-quartile range. It should be noted that for Gaussian population, only 0.7% of the values lies outside these (population)

limits. Any value that equal or less than  $Q_1 - (1.5 \times IQR)$  will be considered as outliers and it should be omitted. Same goes to the value from the dataset which having same or greater than  $Q_3 + (1.5 \times IQR)$  value. For this study, treatment of outliers was performed on only two parameters, namely  $C_{29}/C_{30}\alpha(H)$  and  $C_{31}-C_{35}/C_{30}\alpha(H)$ . Those two parameters were selected for treatment because they have an importance in determining the origin of oil pollution.

PCA analysis was conducted using 16 hopane compounds from 26 surface sediment samples.

The PC can be expressed as:

$$Z_{ij} = a_{i1} X_{1j} + a_{i2} X_{2j} + \dots + a_{im} X_{mj}$$

where  $z$  is the component score,  $a$  is the component loading,  $x$  the measured value of variable,  $i$  is the component number,  $j$  is the sample number, and  $m$  is the total number of variables. Eigen analysis of the sampled data was performed to extract the principal components (PCs) of the measured data using two selection criteria, i.e. the screen plot test and the corrected average eigenvalue. The PCs generated by PCA are sometimes not readily interpreted. Therefore, it is advisable to rotate the PCs by Varimax rotation. Varimax rotations applied on the PCs with Eigenvalues more than 1 are considered significant in order to obtain new groups of variables called Varimax factors (VFs). The number of VFs obtained by Varimax rotations is equal to the number of variables in accordance with common features and can include unobservable, hypothetical, and latent variables. VF coefficients having a correlation greater than 0.75, between 0.75 - 0.50, and between 0.50 - 0.30 are considered as 'strong', 'moderate', and 'weak' significant factor loading respectively. In this study, VF coefficients that show strong significant factor loadings will be discussed.

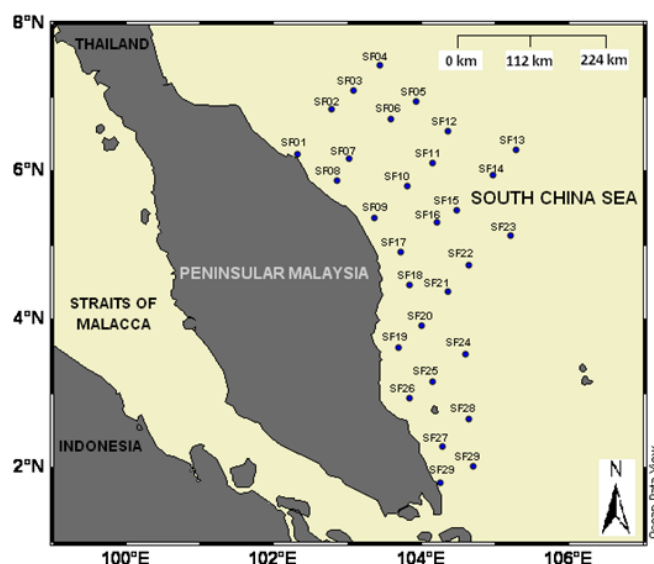


Figure 1: The sampling stations of surface sediment sample in the South China Sea for present study

Table 1: Description of sediment samples in South China Sea collected at various conditions.

Bil	Station code	Longitude (E)	Latitude (N)	Distance (Naut.miles)	Tidal	Weather	Water depth (meter)
1	SF01	102 19.00'	06 13.99'	2.7	High	Cloudy	13
2	SF02	102 47.04'	06 50.04'	50	Low	Cloudy	46.5
3	SF03	103 04.99'	07 05.03'	73	Low	Bright	50
4	SF04	103 26.01'	07 25.98'	100	High	Bright	61
5	SF05	103 56.04'	06 56.09'	108	Low	Bright	52
6	SF06	103 35.17'	06 42.14'	80	High	Cloudy	52
7	SF07	103 01.00'	06 10.00'	40	High	Bright	45
8	SF08	102 51.92'	05 52.10'	15	High	Cloudy	34
9	SF09	103 21.97'	05 22.06'	14	Low	Bright	47
10	SF10	103 48.98'	05 48.20'	48	Low	Bright	55
11	SF11	104 09.11'	06 06.16'	75	Low	Cloudy	72
12	SF12	104 22.11'	06 32.01'	101	Low	Cloudy	59
13	SF13	105 16.99'	06 16.98'	139	High	Cloudy	55
14	SF14	104 58.13'	05 57.15'	115	Low	Cloudy	56
15	SF15	104 29.02'	05 29.08'	80	Low	Cloudy	60.7
16	SF16	104 12.60'	05 18.50'	56	Low	Cloudy	60
17	SF17	103 42.98'	04 54.12'	17	Low	Bright	54
18	SF18	103 49.98'	04 28.14'	20	High	Cloudy	40
19	SF19	103 41.08'	03 37.07'	15	High	Cloudy	23
20	SF20	104 00.05'	03 55.10'	40	Low	Haze	50
21	SF21	104 22.07'	04 22.16'	52	High	Cloudy	65
22	SF22	104 38.44'	04 44.19'	70	High	Cloudy	66
23	SF23	105 12.9'	05 08.10'	109	High	Cloudy	67.2
24	SF24	104 36.00'	03 32.08'	70	Low	Bright	62
25	SF25	104 09.04'	03 09.14'	42	High	Cloudy	41
26	SF26	103 49.97'	02 56.13'	24	High	Bright	20
27	SF27	104 16.97'	02 16.94'	19.5	Low	Haze	30
28	SF28	104 38.91'	02 39.18'	47	High	Cloudy	58
29	SF29	104 41.97'	02 00.55'	35	High	Bright	46
30	SF30	104 15.03'	01 48.04'	4.5	Low	Bright	14

## RESULTS AND DISCUSSION

Figure 2 clearly shows the  $C_{29}/C_{30}$  for surface sediment samples of South China Sea. The ratio ranges between 0.19- 4.00. The maximum ratio for the above station was observed in surface sediment sample collected from station SF22 while the both station of SF09 and SF19 represent the lowest ratio. Figure 3 however shows us the bar chart for  $\Sigma C_{31}-C_{35}/C_{30}$  ratio for the surface sediment samples which ranged from 1.20- 23.32. From the charts, roughly there were four stations (SF02, SF06, SF08 and SF22) possess elevated ratio as compared to other stations. To confirm the extremis of maximum and minimum values in Figure 4.1 and Figure 4.2, an outlier method has been performed on the dataset.

### Data pre-treatment and PCA

The pre-treatment of data prior to PCA can have a significant influence on data outputs, and misleading interpretations may result if pre-treatment is not considered (Reid and Spencer, 2009). *Before further analysis was preceded, 4 surface sediment samples have been neglected due to its extreme value. Those four outliers (SF02, SF06, SF08 and SF22) lied over the third quartile of the outlier treatment analysis.* By conducting an outlier analysis as a tool in statistical analysis (Table 1), calculated  $Q_1-(1.5 \times IQR)$  for  $C_{29}/C_{30}$  and  $\Sigma C_{31}-C_{35}/C_{30}$  were determined as -0.60 and -4.33 for each whereas the value of  $Q_3+(1.5 \times IQR)$  for both ratio were identified as 2.53 and 14.72. The value of  $Q_1-(1.5 \times IQR)$  and  $Q_3+(1.5 \times IQR)$  were then have been applied in order to represent whether or not a generated data is in a state of statistical range. For the purpose outlier analysis, the value of  $Q_1-(1.5 \times IQR)$  has been calculated as lower control limit meanwhile the value of  $Q_3+(1.5 \times IQR)$  was considered as upper control limit (UCL). For the  $C_{29}/C_{30}$  ratio, all points were reliable except the point of SF08 and SF22 where the ratio value (2.73 and 4.00, respectively) were exceeding the UCL (2.53). Meanwhile, for  $\Sigma C_{31}-C_{35}/C_{30}$  value, there were two surface sediment samples that been identified as out of control value ( $> 14.72$ ) where each of them has been collected from the station SF02 and SF06 with their respective extreme value calculated as 16.22 and 86.47. Alternatively, an outlier could be the result of a flaw in the assumed theory, resulting for further calling investigation in this study. However, in this study, Hideshige Takada in his personal communication suggests that an extreme value of  $\Sigma C_{31}-C_{35}/C_{30}$  ratio might due to the natural diagenesis during petroleum production. For example, homohopane distributions are distributed by thermal maturity (Peters and Moldowan, 1991) and hence, the depletion of homohopane ( $\Sigma C_{31}-C_{35}/C_{30}$ ) index might be very high or very low depending on maturity in a suite rocks that derived the oil. So, all extreme values in this study were not reliable to be considered since it represent natural occurring that was happened thousand years ago instead of recent input.

The Eigenvalue and factor loading table from PCA analysis appears on Table 2 and Table 3, respectively. Sixteen variables were analyzed in this study; hence sixteen components have been extracted. From the values appeared in the row, component 1, 2 and 3 demonstrated Eigenvalues of 10.119, 2.423 and 1.378, respectively. Further, only these three components demonstrate Eigenvalues greater than one. This means that 3 components have been retained by the Mineigen criterion. Notice that the first non retained component (component 4) displays an Eigenvalue of approximately 0.975 which of course, is well below 1.0. This is encouraging as the significant level alpha by Pearson correlation was 0.950. The Eigenvalue table also shows that the first three components combined account for approximately 86.989% of the total variance.

Table 2: Outlier analysis on 30 surface sediment samples indicates 4 surface sediment samples were having extreme value of  $C_{29}/C_{30}$  and  $\sum C_{31}-C_{35}/C_{30}$ .

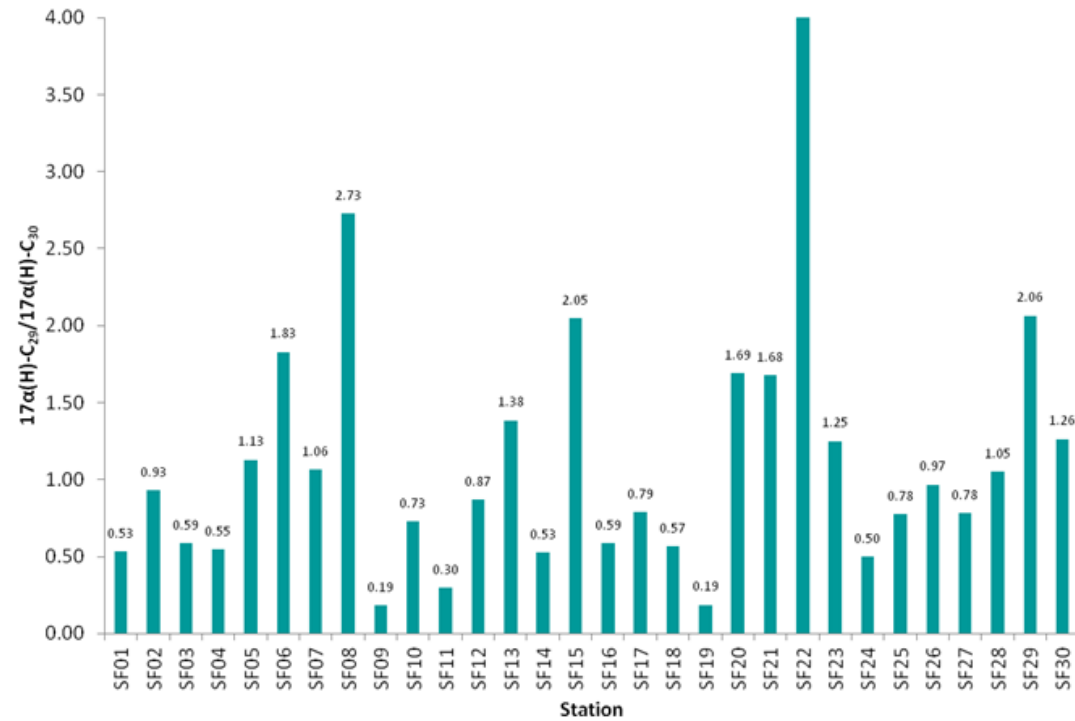
Station	$C_{29}/C_{30}$	$\sum C_{31}-C_{35}/C_{30}$	
SF01	0.53	2.51	Inlying
SF02	0.93	16.22	Outlying
SF03	0.59	8.41	Inlying
SF04	0.55	4.06	Inlying
SF05	1.13	4.42	Inlying
SF06	1.83	86.47	Outlying
SF07	1.06	2.78	Inlying
SF08	2.73	12.07	Outlying
SF09	0.19	2.93	Inlying
SF10	0.73	2.14	Inlying
SF11	0.30	2.40	Inlying
SF12	0.87	4.19	Inlying
SF13	1.38	5.09	Inlying
SF14	0.53	2.22	Inlying
SF15	2.05	9.50	Inlying
SF16	0.59	3.03	Inlying
SF17	0.79	4.56	Inlying
SF18	0.57	3.11	Inlying
SF19	0.19	2.59	Inlying
SF20	1.69	3.00	Inlying
SF21	1.68	3.83	Inlying
SF22	4.00	10.64	Outlying
SF23	1.25	3.76	Inlying
SF24	0.50	1.71	Inlying
SF25	0.78	3.71	Inlying
SF26	0.97	3.34	Inlying
SF27	0.78	9.71	Inlying
SF28	1.05	4.28	Inlying
SF29	2.06	2.76	Inlying
SF30	1.26	14.10	Inlying
Q <sub>1</sub>	0.57	2.82	
Q <sub>3</sub>	1.35	7.58	
IQR	0.78	4.76	
Q <sub>1</sub> -(IQR*1.5)	-0.60	-4.33	
Q <sub>3</sub> +(IQR*1.5)	2.53	14.72	
Outlier $C_{29}/C_{30} = x < -0.60, x > 2.53$			
Outlier $\sum C_{31}-C_{35}/C_{30} = y < -4.33, y > 14.72$			





Table 3: Eigenvalue for each principle component (PC) started to flatten out after third PC(Eigenvalue &lt;1).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Eigenvalue	10.119	2.422	1.378	0.975	0.502	0.231	0.147	0.087	0.055	0.034	0.021	0.013	0.008	0.004	0.002	0.001
Variability (%)	63.241	15.138	8.610	6.094	3.136	1.444	0.920	0.544	0.346	0.215	0.134	0.084	0.050	0.028	0.010	0.007
Cumulative %	63.241	78.379	86.989	93.083	96.219	97.664	98.583	99.127	99.473	99.688	99.822	99.906	99.956	99.984	99.993	100.000

Figure 2: Bar chart of C<sub>29</sub>/C<sub>30</sub> for surface sediment samples

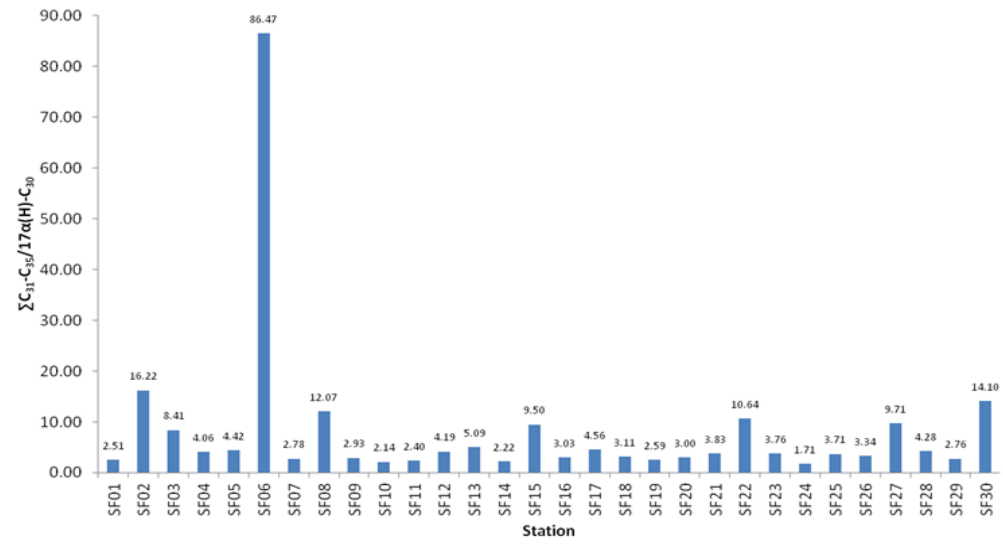


Figure 3: Bar chart of  $\Sigma C_{31}-C_{35}/C_{30}$  ratio for the surface sediment sample

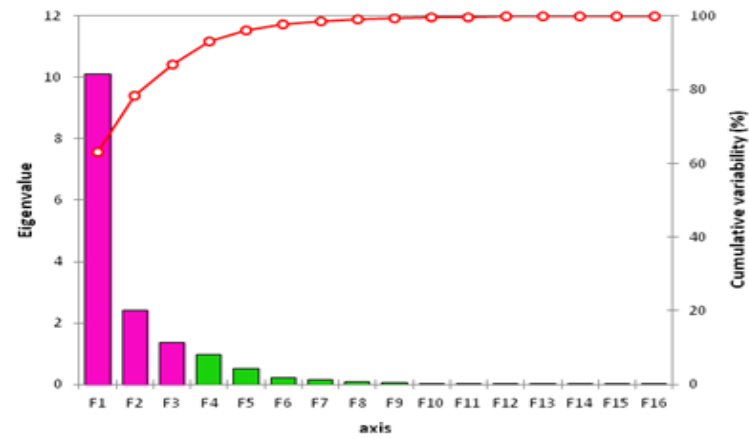


Figure 4: Scree with obvious break after third PC

Table 4: Factor loading for each target compound for 26 surface sediment samples

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Ts	0.829	-0.503	0.113	0.010	-0.155	0.029	-0.076	-0.099	0.014	-0.050	-0.012	-0.017	-0.033	-0.027	-0.006	-0.014
Tm	0.955	-0.123	-0.137	-0.015	-0.130	-0.131	-0.013	-0.125	-0.014	0.010	-0.027	-0.020	0.038	0.011	0.022	-0.001
C <sub>29</sub> - αβ	0.919	-0.049	-0.155	-0.127	-0.024	-0.219	0.250	0.001	0.040	0.002	-0.002	-0.009	0.008	-0.013	-0.016	0.003
C <sub>29</sub> - βα	0.187	-0.128	-0.329	0.905	0.139	-0.034	0.027	-0.015	0.006	-0.016	0.013	0.001	-0.006	0.002	0.000	0.001
C <sub>30</sub> - αβ	0.910	-0.330	-0.112	0.001	0.100	0.053	-0.165	-0.054	-0.015	0.074	-0.015	0.019	0.029	0.001	-0.021	0.006
C <sub>30</sub> - βα	0.631	0.374	0.475	0.236	-0.338	0.221	0.113	-0.068	0.018	0.023	-0.002	0.016	0.001	0.002	-0.001	0.008
C <sub>31</sub> (22S)	0.965	-0.108	0.101	-0.055	0.090	0.067	-0.028	0.044	0.158	-0.042	0.012	-0.013	0.004	0.036	-0.002	-0.004
C <sub>31</sub> (22R)	0.231	-0.027	0.907	0.100	0.294	-0.147	-0.023	-0.031	-0.042	-0.028	0.015	-0.015	0.006	-0.001	-0.001	0.004
C <sub>32</sub> (22S)	0.932	-0.260	-0.101	-0.068	0.064	0.170	0.013	0.079	-0.019	-0.022	0.079	-0.020	0.029	-0.027	0.008	0.004
C <sub>32</sub> (22R)	0.759	0.621	0.009	0.042	-0.023	-0.102	-0.088	0.028	0.036	0.106	0.038	-0.039	-0.031	-0.007	0.004	-0.001
C <sub>33</sub> (22S)	0.927	-0.343	-0.017	-0.061	0.090	-0.017	-0.028	0.062	0.035	-0.005	-0.058	0.028	-0.033	-0.012	0.013	0.018
C <sub>33</sub> (22R)	0.736	0.613	0.113	0.132	-0.121	-0.105	-0.085	0.113	0.008	-0.035	-0.023	0.051	0.026	-0.015	0.000	-0.010
C <sub>34</sub> (22S)	0.872	-0.406	0.144	0.083	-0.034	0.062	0.082	0.138	-0.108	0.043	-0.040	-0.022	-0.007	0.018	0.000	-0.010
C <sub>34</sub> (22R)	0.690	0.653	-0.266	-0.045	-0.050	0.027	-0.068	-0.001	-0.077	-0.087	-0.021	-0.046	-0.007	0.010	-0.009	0.010
C <sub>35</sub> (22S)	0.969	-0.050	-0.113	-0.145	-0.036	-0.075	0.011	-0.045	-0.074	-0.016	0.073	0.059	-0.027	0.023	0.001	-0.001
C <sub>35</sub> (22R)	0.637	0.577	-0.120	-0.120	0.436	0.160	0.096	-0.080	-0.009	0.011	-0.030	0.015	-0.005	-0.007	0.004	-0.009



The Scree plot from this solution appear on Figure 4, where it showed that there were several large breaks in the data following component 1, 2 and 3 then the line begins to flatten out beginning with component 4. The last large break appeared after components 3; this once again suggests that it may be appropriated to retain those three components only.

After extracting the initial components, PCA created an unrotated factor pattern matrix (Table 4) where the row represent the variables been analyzed, and the column represent the retained components. The entries in the matrix are factor loading in which it is a general term for a coefficient that appears in a factor matrix where the rows for observed variables intersect with the column for PC1 For example, correlation between Ts (first variable) and the first component is 1; the correlation between Tm (second variable) and the first component was 0.867, and so forth.

In order to determine the item construct seem to be measured by component 1 and so on, correlation between the variable need to be reviewed before further interpretation can be conducted. Unfortunately, when more than one component has been retained in an analysis, the interpretation of unrotated factor is usually quite difficult (Table 3). To make interpretation easier, rotation has been performed (Table 4). Varimax rotation is an orthogonal rotation in which it tends to maximize the variance of a column. Interpreting a rotated solution involves identifying the variables that demonstrate high loadings for a given component. The first decision to be made at this stage is to decide how large a factor loading must be to be considered 'large'. For the purpose of this study, a loading considered as large as its absolute value exceeds 0.750. All meaningful loading (loading $\geq$ 0.750) have been highlighted (Table 4).

From Table 4, we can say that factor 1 (PC1) was influenced Ts, Tm, 17 $\alpha$ (H), 21 $\beta$ (H)-norhopane, 17 $\alpha$ (H), 21 $\beta$ (H)- hopane, 22S C<sub>31</sub>-C<sub>35</sub> compounds. The correlation between all of them was closely related to the variable with the value of correlation was greater than 0.80. Since Ts and Tm also related to the thermal maturity (Zakaria 2000), the 22S C<sub>31</sub>-C<sub>35</sub> homohopane indices will be fluctuated with decreasing and increasing amount of Ts and Tm. Those components considered as excellent in tracing both oil pollution and variation of anthropogenic input to sediment (Cranwell and Koul, 1989; Uemura and Ishiwatari, 1992). Both Ts and Tm are having degrades less during diagenesis and catagenesis where they are generally unaffected by even severe biodegradation. If biodegraded, these two isomers are removed at approximately the same rate, preserving their initial ratio (Peters et al., 1993). For example, severely biodegraded oils from the Phosporia formation in Wyoming have altered the distribution of steranes, hopanes and mono and triaromatic steroid, but Ts and Tm remain unaltered (Sundararaman and Hwang, 1993). A part of that, the ratio of Ts and Tm also can be used as indicator of thermal maturity when comparing oil or sediment sample from the same source (Volkman et al., 1997). Thus the ratio of Tm to Ts is an indicator of the maturity of a rock due to the less sensitivity on thermal condition when capering oil or sediment samples from the same source (Booyatumanond et al., 2006). The 17 $\alpha$ (H), 21 $\beta$ (H)- norhopane, 17 $\alpha$ (H), 21 $\beta$ (H)- hopane and 22S, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane have been found from bacteria that derived the petroleum, hence can be abundantly found in the sediment (Uemura and Ishiwatari, 1992). The 17 $\alpha$ (H),21 $\beta$ (H) of C<sub>27</sub>- C<sub>35</sub> stereochemistry range predominate in indicating a substantial contribution from petroleum, where they are giving characteristic to petroleum because of their greater thermodynamic stability compared to the other epimeric ( $\beta\beta$ ,  $\beta\alpha$ ) series. The 17 $\alpha$ (H),21 $\beta$ (H) of C<sub>27</sub>- C<sub>35</sub> derived from bacteria that derived from petroleum (Uemura and Ishiwatari, 1992) and also

can be formed by diagenetic acid catalyzed isomerization or microbial processes (Rohmer et al., 1980; Brassell et al., 1980). Moreover, they are also abundantly found in sediment (Uemura and Ishiwatari, 1992) which in mature sample. The 17 $\alpha$ (H), 21 $\beta$ (H)- stereoisomers are greatly predominate over 17 $\beta$ (H), 21 $\alpha$ (H)- stereoisomers, so that the concentration of 17 $\alpha$ (H), 21 $\beta$ (H)- stereoisomer is more significant in crudes. High abundance of the 17 $\alpha$ (H), 21 $\beta$ (H)- norhopane is also associated with oils derived from carbonate source rock (Volkman et al., 1997). A part of that, the 17 $\alpha$ (H), 21 $\beta$ (H)- hopane is neither generated nor degraded in laboratory experiments that mimic biodegradation in the field, so hopane does have the appropriate properties to serve as a conserved internal marker. In other manner, the 17 $\beta$ (H), 21 $\alpha$ (H)- hopane derivatives was generated in sediments at the earliest diagenetic stage (Quirk et al., 1984) would cause decrease in abundance after moderate maturation and this left 17 $\alpha$ (H), 21 $\beta$ (H)- hopane dominant (Mackenzie, 1984) in concentration compare to 17 $\beta$ (H), 21 $\alpha$ (H)- hopane.

The correlation circle (Figure 5) interprets the circle correlation of the factor loading after Varimax rotation performed on it whereby it is useful in interpreting the meaning of the axes. The vertical axis in Figure 2 was linked with individual hopanes from PC1 which explained 48.82% contribution in determining the source of oil pollution as per described in previous paragraph. Horizontal axis has been contributed by the rest of hopanes which do not include in vertical axis (PC2=28.82%) such as of 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- bishomohopane, 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- trishomohopane, 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- tetraomohopane and 22R, 17 $\alpha$ (H) and 21 $\beta$ (H)- pentakishomohopane except 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane.

Relatively, the compounds in PC2 (Table 4 and Figure 4) were less reliable to be used as compared to the 22S configuration for tracing oil pollution because the former concentration was generally very low (Uemura, 1992). As been proposed by Rohmer et al. and Brasell et al. in 1980, those compounds derived from bacteria peat and higher plant where they had been formed by diagenetic acid catalyzed isomerization and or microbial process. For example, they were formed naturally during decay of plant material (Quirk et al., 1984) and became major constituent of Windmere River (Cranwell and Koul, 1989).

The third component (PC3) was heavily loaded with 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane (factor loading=0.962). The 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane derived from bacteria associated with peat and or higher plant (Quirk et al., 1980; Vandorsselaer et al., 1977). Under certain cases, the abundance of this compound can be greater than 22S configuration where it is effective to be used in tracing oil pollution sources (Uemura and Ishiwatari, 1992). However, in this study the concentration of 22R, 17 $\alpha$ (H), 21 $\beta$ (H)- homohopane give less significant impact in determining the source in which the percentage of the component plus 17 $\beta$ (H), 21 $\alpha$ (H)- norhopane and 17 $\beta$ (H), 21 $\alpha$ (H)- hopane was given by 22.36%. Hence, this compound in this study less reliable to be used for identifying oil pollution sources.

Table 5: Factor loading after Varimax rotation.

	PC1	PC2	PC3
Ts	0.958	0.010	0.189
Tm	0.877	0.420	-0.030
C <sub>29</sub> - $\alpha\beta$	0.808	0.465	-0.049
C <sub>29</sub> - $\beta\alpha$	0.251	0.024	-0.310
C <sub>30</sub> - $\alpha\beta$	0.949	0.220	-0.019
C <sub>30</sub> - $\beta\alpha$	0.289	0.606	0.558
C <sub>31</sub> (22S)	0.859	0.416	0.207
C <sub>31</sub> (22R)	0.139	0.015	0.926
C <sub>32</sub> (22S)	0.929	0.289	-0.002
C <sub>32</sub> (22R)	0.299	0.926	0.121
C <sub>33</sub> (22S)	0.963	0.209	0.077
C <sub>33</sub> (22R)	0.276	0.898	0.221
C <sub>34</sub> (22S)	0.939	0.111	0.229
C <sub>34</sub> (22R)	0.244	0.943	-0.159
C <sub>35</sub> (22S)	0.847	0.487	-0.002
C <sub>35</sub> (22R)	0.230	0.836	-0.024

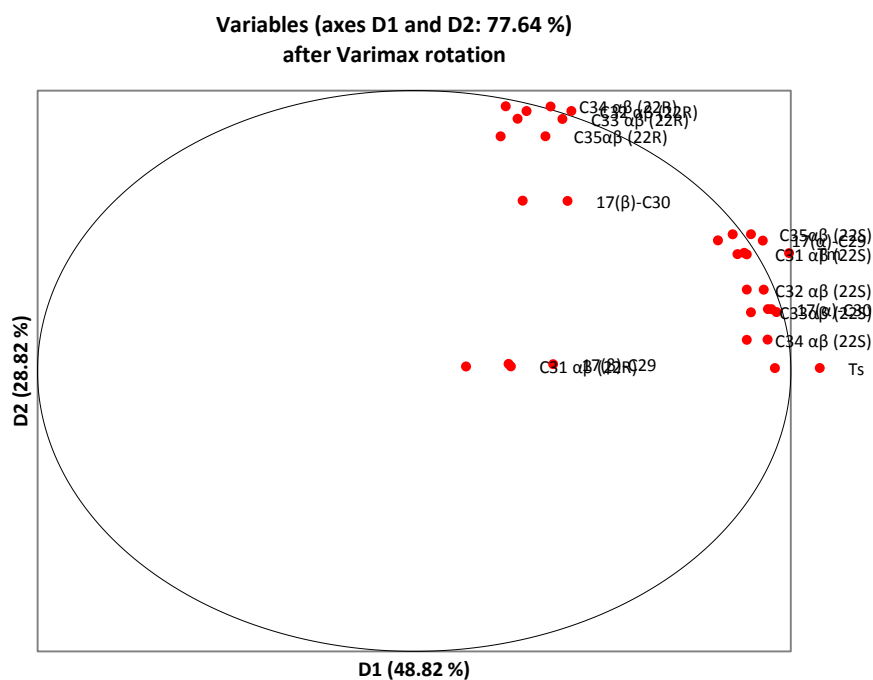


Figure 5: Circle correlation of factor loading after Varimax rotation



## CONCLUSION

By using an outlier we found that surface sediments samples from four stations (SF02, SF06, SF08 and SF22) were not reliable to be included. So, out of 30 surface sediment samples collected from Southwest corner of South China Sea, only 26 surface sediments will be considered for further analysis.

A part of that, from the description mentioned, proving that the nine individuals hopane from PC1 have the interests of its own in determining the source of oil pollution. Their chemical properties are stable in terms of structure and thermal as described makes them more dominant than the other individual hopanes. From PCA analysis, some problems in understanding the selectivity of certain individual hopane to determine the origin and sources of petroleum pollution can be explained. If observed carefully, most researchers use the ratio of  $C_{29}/C_{30}$  to determine the origin of a sample. Both ratios only take into account the concentration of  $17\alpha(H),21\beta(H)$ -stereoisomers, while the concentrations  $17\beta(H),21\alpha(H)$ -stereoisomers for  $C_{29}$  and  $C_{30}$  are not necessarily to be included in the determination of the ratio. This question has been answered by verification analysis PCA, which, as explained  $17\alpha(H),21\beta(H)$ -stereoisomers has a high dominance over  $17\beta(H),21\alpha(H)$ -stereoisomers. In other words,  $17\beta(H),21\alpha(H)$ -stereoisomers did not give a significant effect if the concentration is not taken into account in calculating the ratio.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge financial support of Ministry of Science, Technology and Innovation, Malaysia (MOSTI) who provided partial funding for this project (Science Fund Project No. 5450095). Dr. Azadeh Shahbazi and Mr. Asamudin Hasan who helped with the sampling and reviewing the manuscript respectively are kindly acknowledge.

## REFERENCES

- Asuquo, F.M. 1991. Tar-balls on Ibeno–Okposo beach of South-East Nigeria. *Mar Pollut Bull*, **22**, 150–151.
- Boonyatumanond, R., Wattayakorn, G., Togo, A., Takada, H., 2006. Distribution and origins of polycyclic aromatic hydrocarbons (PAHs) in riverine, estuarine, and marine sediments in Thailand. *Mar. Pollut. Bull.* **52**, 942-956.
- Brasell, S.C., Comet, P.A., Eglinton, G., Isaacson, P.J., McEvoy, J., Maxwell, J.r., Thomson, I.D., Tibbetts, P.J.C, Volkman, J.K., 1980. The origin of lipids in the Japan Trench. A.G. Douglas, J.R. Maxwell (Eds), *Advance in Organic Geochemistry 1979*, Pergamon, Oxford, 375-392.
- Buncher, C.R., Tsay, J.Y., 2006. *Statistics in the pharmaceutical industry*; CRC Press: US.
- Chandru, K., Zakaria, M.P., Anita, S., Shahbazi, A., Askari, M., Bahry, P.S., Mohamed, C.A., 2008. Characterization of alkanes hopanes and polycyclic aromatic hydrocarbon (PAHs)

- collected from the East Coast of Peninsular Malaysia. *Mar. Pollut. Bull.* **56**,950-962.
- Cranwell, P.A., Koul, V.K., 1989. Sedimentary record of polycyclic aromatic and aliphatic hydrocarbon in the Windermere catchment. *Water Research* **2**, 275-283.
- Hawkins, D.M., 1980. Identification of outliers: Monographs on applied probability and statistics; Chapman and Hall Ltd: New York.
- 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(1)**, 26-42.
- Mackenzie, A.S., 1984. Application of biological markers in petroleum geochemistry. Brooks, ., Welte, D. (Eds.), *Advance in petroleum geochemistry*, Vol 1 Academic Press, London, 115-214
- Morton, B. and Blackmore, G., 2001. South China Sea. *Mar. Pollut. Bull.* **42**,1235-1263
- Peters, K.E. and Moldowan, J.M., 1993. *The biomarker guide: Interpreting molecular fossils in petroleum and ancient*. N.J Prentice-Hall: Eaglewood Cliffs.
- Peters, K.E., Walters, C.C., Moldowan, J.M., 2005. *The biomarker guide: Volume 1 biomarkers and isotopes in the environment and human history*, Second edition, pp 471. Cambridge: University Press, NJ.
- Quirk, M.M., Patience, R.L., Maxwell, J.R., Wheatley, R.E., 1980. Recognition of the sources of isoprenoid alkanes in recent sediments. In *Analytical technique in Environmental Chemistry*, ed. J. Albaiges, pp 23-31. Oxford: Pergamon.
- Quirk, M.M., Wardroper, A.M.K, Wheatley, R.E., Maxwell, J.R., 1984. extended hopanoid in peat environments. *Chem. Geol.* **42**, 25-43.
- Reid, M.K., Spencer, K.L., 2009. Use of principal components analysis (PCA) datasets: The effect of data pre-treatment. *Environmental Pollution* **157**, 2275–2281
- Rohmer, R., dastillung, M., Ourisson, G., 1980. *Naturwissenschaften* **67**,456-458.
- Sundararaman, P., Hwang, R.J., 1993. Effect of biodegradation on vanadylporphyrin distribution. *Geochim. Cosmochim. Acta* **57**, 1367-1377.
- Uemura, H., Ishiwatari, R., 1992. Aliphatic hydrocarbons and hopanes in lake Haruna sediments with special reference to oil pollution. *Jap. J. limnol.* **53**, 293-303
- U.S. National Academy of Sciences: Oil in the sea III: Inputs, fates and effects. Report 2002by the National Research Council (NRC) Committee on Oil in the Sea: Inputs, Fates, and Effects.
- Vandorsselaer, A., Albrecht, P., Ourisson, G., 1977. Odentification of novel 17 $\alpha$ (H)- hopanes in shales, coals, lignites, sediments and petroleum. *Bull. Soc. Chim.* 165-176.
- Volkman, J.K., Revill, A.T. and Murray, A.P., 1997. Applications of biomarkers for identifying sources of natural and pollutant hydrocarbons in aquatic environments. In *Molecular Markers in Environmental Geochemistry*, ed. Eganhouse, R.P. pp. 83-99. Washington DC: ACS Symposium Series.
- Wang, Z., Stout, S. and Fingas, M., 2006. Forensic Fingerprinting of Biomarkers for Oil Spill Characterization and Source Identification. *Environmental Forensics* **42(7)**, 105-146.
- Wang, Z., Yang, C., Fingas, M., Hollebone, B., Yim, U.H., Oh, J.R., 2007. Petroleum biomarker fingerprinting for oil spill characterization and source identification. In *Oil Spill Environmental Forensics*, ed. Wang, Z. and Stout, S.A., pp. 73-14. London: Academic Press.
- Waples, D.W. and Machihara, T., 1991. Biomarkers for geologists-a practical guide to the application of steranes and triterpanes in petroleum geology. *Amer.*

*Assoc. Petrol. Geol., Methods in Exploration*, No. 9.

- Zakaria, M.P., Okuda, T., Takada, H., 2001. Polycyclic aromatic hydrocarbon (PAHs) and hopanes in stranded tar-balls on the coasts of Peninsular Malaysia: applications of biomarkers for identifying sources of oil pollution. *Mar. Pollut. Bull.* **42**, 1357–1366.
- Zakaria, M.P., 2001. Distribution, sources and transport pathway of petroleum hydrocarbons in Malaysia rivers, estuaries and the Straits of Malacca. Ph.D Dissertation, Tokyo University of Agriculture and Technology, Japan
- Zakaria, M.P., Takada, H., Tsutsumi, H., Ohno, k., Yamada, J., Kouno, E. and Kumata, H. 2002. Distribution of polycyclic aromatic hydrocarbons (PAHs) in rivers and estuaries in Malaysia; a widespread input of petrogenic PAHs. *Environ. Sci. Technol.* **36**, 1907-1918.