Classification of weathered petroleum oils by multi-way analysis of gas chromatography-mass spectrometry data using PARAFAC2 parallel factor analysis

Author/Contributor:
Ebrahimi Mohammadi, Diako; Li, Jianfeng; Hibbert, D. Brynn

Publication details:
Journal of Chromatography A
v. 1166
Chapter No. 1-2
pp. 163-170
0021-9673 (ISSN)

Publication Date:
2007

Publisher DOI:
http://dx.doi.org/10.1016/j.chroma.2007.07.085

Classification of weathered petroleum oils by multi-way analysis of gas chromatography-mass spectrometry data using parallel factor analysis (PARAFAC2)

Diako Ebrahimi, Jianfeng Li and David Brynn Hibbert*

School of Chemistry, University of New South Wales, Sydney, NSW 2052, Australia

*Corresponding author

Tel: +61 2 9385 4713; Fax: +61 2 9385 6141; Email: b.hibbert@unsw.edu.au

Abstract

The application of multi-way parallel factor analysis (PARAFAC2) is described for the classification of different kinds of petroleum oils using GC-MS. Oils were subjected to controlled weathering for 2, 7 and 15 days and PARAFAC2 was applied to the three-way GC-MS data set (MS × GC × sample). The classification patterns visualized in scores plots and it was shown that fitting multi-way PARAFAC2 model to the natural three-way structure of GC-MS data can lead to the successful classification of weathered oils. The shift of chromatographic peaks was tackled using the specific structure of PARAFAC2 model. A new preprocessing of spectra followed by a novel use of analysis of variance (ANOVA)-least significant difference (LSD) variable selection method were proposed as a supervised pattern recognition tool to improve classification among the highly similar diesel oils. This lead to the identification of diagnostic compounds in the studied diesel oil samples.

1. Introduction

Gas chromatography-mass spectrometry (GC-MS) used in the analysis of petroleum oils [1,2] provides an extracted ion chromatogram (EIC) for a number of prescribed *m/z* channels. Each EIC expresses hydrocarbon constructional isomers having the same *m/z*. Petroleum oils are distinguished based on the differences of patterns (known as 'hydrocarbon fingerprints') they exhibit in their EICs. Among thousands of different compounds that exist in oils, those which are "source specific" and "weathering stable" are used for identification of source of oil spills [1-4]. Two methods, published by the American Society for Testing and Materials (ASTM D 5739-00) [2] and the Nordisk Innovations Centre (Nordtest) [4,5] are the major approaches for source identification of oil spills by GC-MS. In the ASTM method, matching the oil spill to the suspected source(s) is through overlaying and visually comparing the EICs. With Nordtest methodology a number of diagnostic ratios (DRs, ratios of chosen isomer peak

areas) of the oil spill spectra are plotted against those from each suspected source. A conclusion is then reached based on the approach of DR points to the straight-line (spill ratio = suspect ratio) within their measurement uncertainties [4]. Both methods must contend with possible degradation of a spilled oil by exposure to the environment. Important mechanisms of degradation include bacterial action and photo-oxidation. It has been shown that bacteria degrade some isomers in preference to others; for example with methyldibenzothiophenes (MDBTs), the isomers 2- and 3-MDBT are preferred by bacteria above others and 4-MDBT is most stable with respect to this mechanism; [4,6] or in the case of methylphenanthrenes (MPhs), the greater photo-oxidation of 1-methylphenanthrene (1-MPh) over 2-MPh has been reported by different groups [7,8]. Different weathering mechanisms will cause spectral discrepancies in the EICs and might be misleading in conventional visual comparison of chromatograms. Moreover with visual comparison of spectra there could be a risk of subjective errors,[9] particularly in the cases that are not clear-cut, such as comparisons of highly similar diesel oils. The effect of biodegradation is kept to a minimum in the Nordtest methodology by exclusion of the peaks of those isomers preferred by the bacteria. However photo-oxidation [7,8,10] is not considered in assigning DRs, perhaps because the majority of the DRs calculated in the Nordtest approach have been first offered by geochemists as biomarkers for petroleum exploration where there is a greater interest in long term weathering mechanisms, such as biodegradation. Compared to biodegradation which is the last fate of petroleum oil in the environment, [10] photo-oxidation is short term and its depletion effect on hydrocarbons is opposite to that in biodegradation. It has been shown that bacteria target 2- and 3-MP more than 9- and 1-MP, while the latter two exhibit more sensitivity to photo-oxidation [7,8] or when MDBTs are concerned, the 4-MDBT is the most stable in terms of biodegradation and least stable in terms of photo-oxidation [6,7]. This could make some of the proposed PAH (polyaromatic hydrocarbon) diagnostic ratios such as 4-MDBT/1-MDBT less informative in photo-oxidized oil spills. It has also been reported that while biodegradation of PAHs is less with increasing ring numbers and branches, photo-oxidation is greater [8,10]. Therefore the use of other PAH diagnostic ratios such as C2-DBT/C2-Ph, C3-DBT/C3-Ph, C3-DBT/C3-Chr and Retene/C4-Ph could also be misleading in photo-oxidised oils. (Prefixes C2 to C4 indicate the number of substituted carbons; Ph and Chr are abbreviations for phenanthrene and chrysene, respectively). It is apparent from Prince *et al*'s work [10] that those ratios do not remain unchanged during the photo-oxidation process. An alternative would be the use of resistant biomarkers (such as hopanes, steranes and triaromatic steroids) to form diagnostic ratios [1,2,4]. They have long been utilized in assigning crude oil (or other heavy oil) spills to their responsible sources, but issues arise when lighter and refined petroleum products such as diesel oils are concerned. Many of these stable diagnostic molecules are removed (or remain in a low and so non-measurable concentration) from the oil during refining [11]. In contrast to the ASTM method, successful assessment in the Nordtest methodology requires reasonable separation of isomer chromatographic peaks and knowledge of the origin of each peak in an EIC.

Identification of various biodegraded heavy oils using the Nordtest method has been extensively reported [1], and therefore is not reiterated in this paper. In this study a number of oil samples were weathered using a simple experiment to ensure that the weathering process was dominated by photo-oxidation (and evaporation) processes, and then an objective chemometrics methodology was applied to classify the weathered oils. The reported method does not require the peaks to be separated or identified and all the isomers (i.e. the entire EIC) are included in the analysis. It is demonstrated that exploratory analysis of oil samples is possible by fitting a suitable multi-way model (PARAFAC2) to GC-MS data. Supervised pattern recognition using a proposed analysis of variance (ANOVA) – least significant

difference (LSD) variable selection applied to the baseline corrected and smoothed EICs is shown to improve the discrimination among very similar diesel oils.

2. Methodology

2.1. Exploratory data analysis

A GC-MS analysis of a typical oil sample provides an array of data of size (number of samples × elution time points × number of *m/z* channels). This three-way structure (samples × elution times × *m/z* channels) of GC-MS data implies that fitting a proper three-way model would be more successful than an unfolded two-way principal component analysis (PCA) in which one mode (e.g. *m/z*) is coalesced with another (e.g. elution time) [12]. In PARAFAC [13] the data cube is decomposed into one scores and two loadings matrices representing sample (concentration), chromatographic and *m/z* profiles, respectively. Classification using exploratory data analysis is one of the applications of PARAFAC [13,14] which can be applied to any type of data within which a trilinear (three-way) structure is preserved [14]. In chromatography, and so GC-MS, the trilinear structure within data is destroyed if chromatographic peaks shift in time from run to run (sample to sample) [14-16]. A modification of PARAFAC known as PARAFAC2 [15,16] has been reported that can model data in presence of such shifts. The theory and algorithms for fitting PCA, PARAFAC and PARAFAC2 models have already been discussed by many authors [13-16] and are not reiterated here, but a graphical presentation of the methods is given in Figure 1. It is worth noting here that an alternative approach in this work would be to apply PARAFAC to the aligned chromatograms by means of a chromatographic shift correction method [17] such as correlation optimized warping (original [18] or modified [19] versions) and reduced set mapping [20]. PARAFAC2 was used here to avoid the computational efforts of the alignment step and so to simplify the procedure. Moreover, as will be discussed later in this paper, the weathering experiment significantly alters the composition of some of the extracted ion chromatograms such as naphthalenes that are depleted up to 100%. This could potentially pose a challenge to the alignment methods designed for more or less similar samples [21].

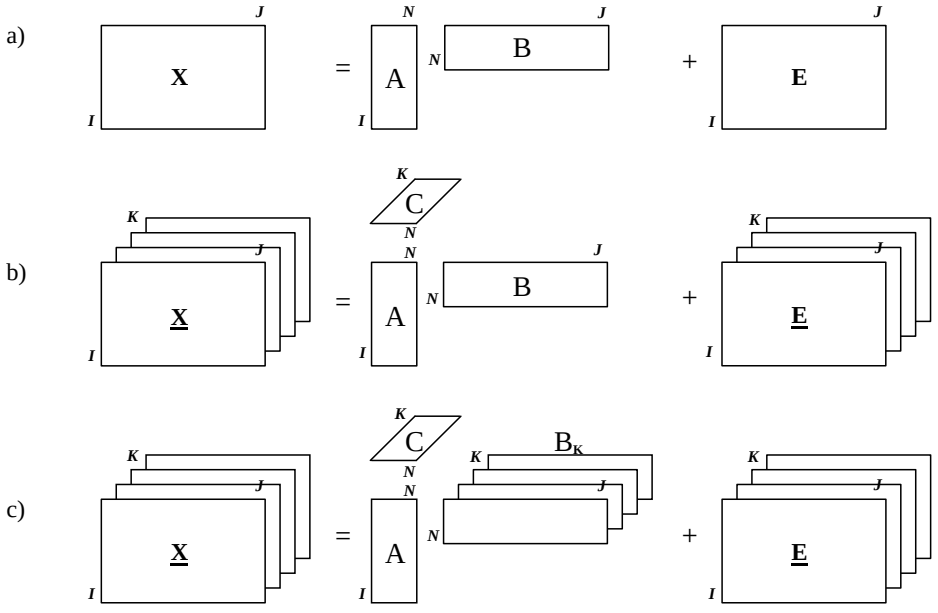

Figure 1. Graphical presentation of three multivariate approaches of a) PCA, b) PARAFAC and c) PARAFAC2.

In Figure 1 for PARAFAC (1b) and PARAFAC2 (1c), the subscript *K*, number of frontal slabs, represents the third mode, **C**. $\mathbf{X}_K$ is the *K*th frontal slab of three-way array $\underline{\mathbf{X}}$. **A** and **B** are matrices holding the profiles in the first and second modes respectively. The difference between PARAFAC and PARAFAC2 is in **B**, the second mode profile. The subscript *K* given to **B** in PARAFAC2 implies that there are *K* individual loading matrices ($\mathbf{B}_1$-$\mathbf{B}_K$), one for each frontal slab, while there is only one common **B** matrix in PARAFAC. It demonstrates that PARAFAC2 allows for certain profile variability of one of the modes with one of the other two modes and will able to model this type of data, where PARAFAC is likely to fail. GC-MS data in which sample-to-sample chromatographic shifting occur are perfect candidates for PARAFAC2. Suppose 10 samples are analyzed by GC-MS and the data are recorded for 100 *m/z* channels in 600 seconds (10 min, 1 second interval). The data can be arranged into an array of size 100 × 600 × 10, displaying 100 rows (*m/z* channels), 600 columns (elution time points) and 10 samples. This is the proper arrangement of GC-MS data to be modeled by PARAFAC2 in which the second mode profiles (chromatograms) change with the third mode (samples). Subsequent to fitting the proper PARAFAC2 model to the data, classification of samples can be achieved by visualization of the results in the samples profile (scores) plots.

2.2. Supervised pattern recognition with variable selection by ANOVA-LSD

In exploratory data analysis no prior information about the samples is considered in the classification [22]. This approach is in contrast to a trained model optimized with known samples to ensure the most effective classification. To supervise the classification ANOVA [23] can be used as a variable selection tool [24-26] by which the variables that can distinguish between different instances of a factor are selected. In the present case the factor is 'oil sample' comprising different oil samples as its instances. The fresh and weathered members of a particular oil are considered as repeated observations of each instance of the factor. Thus a one-way ANOVA table is formed for each *m/z* channel (variable) giving an output of the variance between the instances of the factor, and the residual (within instance) variance. Any *m/z* which gives a significant (P(*f*>F) < 0.05) *F* value (as the ratio of between instance variance and within instance variance [23]) is a potentially diagnostic variable. To select the diagnostic variables amongst potentially diagnostic variables, we propose the use of the method of least significant difference (LSD [23]) in a pairwise comparison of the means of instances for each potentially diagnostic *m*/*z* (variable). LSD is calculated by

$$LSD = t_{\alpha,df} \times s_{wi} \times \sqrt{\frac{2}{n}} \qquad (1)$$

where $t_{\alpha,df}$ is the two tailed Student's *t* value at probability α and *df* within instance degrees of freedom. s_{wi} is the within instance standard deviation and *n* is the number of observations in each instance of the factor (here 4 = fresh plus three weathered oils). Any difference between the means of two instances that is greater than LSD is significant at a probability α. A 'classification percentage' of a *m/z* is defined as the number of significant differences between the means of instances divided by the total number of mutual comparisons, expressed as a percentage. The diagnostic variables are selected by setting an optimum threshold on the classification percentage. An example layout of the proposed variable selection method using seven *m/z* and four different oils is given in Figure 2.

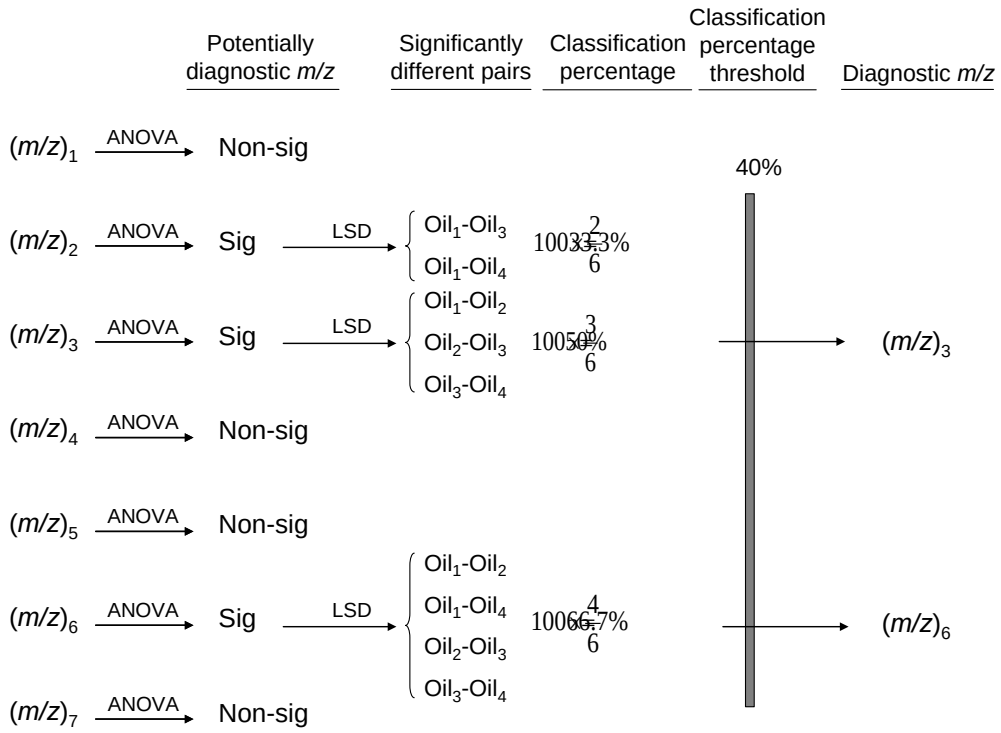

Figure 2. Typical layout of the proposed ANOVA based variable selection method using seven *m/z* and four different oils. Sig: Significant; Non-sig: Non-significant.

As displayed seven ANOVAs are employed, one for each *m/z*. Three *m/z*, namely $(m/z)_2$, $(m/z)_3$ and $(m/z)_6$ out of seven are found to be potentially diagnostic by ANOVA. LSD is then applied to these three *m/z*. There are four oil groups and therefore six possible mutual comparisons of oil_1-oil_2, oil_1-oil_3, oil_1-oil_4, oil_2-oil_3, oil_2-oil_4 and oil_3-oil_4. The significantly different pairs (those for which the difference between their means is greater than LSD) are shown for each potentially diagnostic *m/z*. The classification percentage is calculated by dividing the number of significantly different pairs by six then multiplied by 100. By setting a threshold of 40% (which was set for the data here) only $(m/z)_3$ and $(m/z)_6$ are chosen as diagnostic.

3. Experimental

3.1. Oil samples

A set of 17 different petroleum oils, ranging from transformer and lubricating oils to crude and diesel oils, as well as mixtures of them, were obtained from the New South Wales Department of Environment and Conservation (NSWDEC). A list of all studied oil samples with their kind and origin is given in Table 1.

Table 1: Oil samples studied, with their kinds and origins

Code	Type	Origin
D1	Diesel	Retail outlet, Sydney, NSW, Australia
D2	Diesel	Retail outlet #1, Mudgee, NSW, Australia
D3	Diesel	Suspect source oil from terrestrial oil spill, Orange, NSW, Australia
D4	Diesel	Suspect source oil from marine oil spill, Nelsons Bay, NSW, Australia
D5	Diesel	Retail outlet #2, Mudgee, NSW, Australia
D6	Diesel	Suspect source oil for terrestrial oil spill, Lithgow, NSW, Australia
D7	Diesel	Suspect source oil from terrestrial oil spill, Lithgow, NSW, Australia
D8	Diesel	Retail outlet #3, Mudgee, NSW, Australia
DR	Diesel	EPA reference #2 Fuel oil
CA	Crude	Gippsland, Victoria, Australia
CM	Crude	Tapis, Malaysia
CU	Crude	EPA reference, Louisiana, USA (suspected of being a miss-labelled diesel – see text for discussion)
CK	Crude	EPA reference, Kuwait
T	Transformer	Shell Diala S, Australia
L	Lubricating	Pennzoil HD SAE 30, Victoria, Australia
DL1	Mixture of diesel and lubricating	Recovered oil from spill incident, Sydney, NSW, Australia
DL2	Mixture of diesel and lubricating	Recovered oil from marine oil spill, Sydney, NSW, Australia

3.2. Weathering procedure

Oil samples were subjected to a regime of weathering by placing a 2-5 mm thick slick of each oil over water in a beaker, which was then exposed on the roof of a building for 2, 7 and 15 days from 21 November to 4 December 2002 (for similar methods refer to references 7 and 8). During this time in Sydney, Australia the weather was dry and hot, with three days having a maximum temperature over 35 °C.

3.3. Chemical analysis

Oil samples were analysed by a Varian 2000 ion trap GC-MS instrument. A 30 m × 0.25 mm, 0.25 μm DB-1 Phenomenex GC column was used. The injected volume of 2 μL in splitless injection mode (1 minute splitless time) and injection temperature of 280 °C were applied. The initial GC oven temperature was set to 40°C for 4 min, followed by an increase to 300 °C at the rate of 10 °C min^{-1} and held for 10 min. 1 mL min^{-1} helium flow was used constantly and masses were scanned from *m/z* 108 to 230 between 11 to 45 minutes (points taken at 0.003 min intervals) from the start of the run. Consequently 123 *m/z* were scanned for nearly 11336 time points for 68 (17 × 4) oil samples. Thus the data represents 123 EICs for 68 samples.

3.4. Data analysis

3.4.1. Exploratory data analysis

No significant peak was found in the last 5 min (40 to 45 min after injection) of EICs, so this part was removed from the data set and 9670 elution time data points (11-40 min, 0.003 min interval) remained for further analysis. Then, a set of 37 target *m/z* out of the 123 were chosen as likely to be highly discriminating and diagnostic among the oil types. Those 37 *m/z*

have already been identified as associated with molecules such as naphthalenes, phenanthrenes, anthracenes, dibenzothiophenes, steranes, terpanes and some other molecules that can be used to discriminate oils by ASTM, Nordtest and other oil spill identification methods [1,2,4]. No priority is given to a particular *m/z* and any *m/z* already known as diagnostic in the literature[1,2,4] has been included in this work. Thus the new data array comprised 37 EICs (9670 time points in each) recorded for 68 samples. The list of all 37 characteristic ions used in this paper with their corresponding target compounds is given in Table S1 of the Supplementary Information. The 37 EICs of 68 samples were arranged in an array of size 37 × 9670 × 68 which was subjected to decomposition by PARAFAC2. The resultant cube of data was mean-centred across all three modes and scaled within its first and third modes [14] before applying the PARAFAC2 program.

3.4.2. Supervised pattern recognition with variable selection by ANOVA

To improve the classification, supervised pattern recognition using ANOVA variable selection [24-26] was applied to a sub-set of data from similar diesel oils. A data array including local diesel samples D2-D8 was constructed having 28 samples (D2 to D8 each with 4 members) in its third mode, 9670 elution time points in its second mode and 37 *m/z* channels in its first mode. The row vectors (EICs) of this cube of data were smoothed using a second order polynomial Savitsky-Golay filter with fifteen data points in each [27]; the first derivative of the data was then calculated. This was followed by a second smoothing step to reduce the noise raised due to derivatization. EICs were then normalized to 100 to remove the variability due to the concentration difference of samples and to bring the samples (and EICs) into a comparable scale. To show the effect of this processing method a raw EIC (*m/z* 190, C4-benzo[b]thiophenes) along with its derivative and smoothed spectra is shown in Figure S1 of the Supplementary Information. The applied pre-processing also improves the classification of weathered diesel oils which exhibit considerable noise and baseline drift. To prepare the ANOVA tables it was necessary to represent each EIC vector (*m/z*) by a single value. The Frobenius norm was chosen in this work to ensure the shift in the position of the chromatographic peaks (even non-linear shifts) are not likely to have an impact on the variable selection, as all peaks contribute to the Frobenius norm, irrespective of position. Note that because each EIC is represented by its Frobenius norm, using derivative and smoothed spectra is a prerequisite when the oil samples within a group exhibit different baseline and noise levels. For ANOVA variable selection of the diagnostic *m/z*, the Frobenius norm of each sample of each oil was calculated for each EIC (*m/z*). The resultant data were recorded in 37 matrices of size 4 (fresh plus three weathered samples of an oil) × 7 (oils D2 – D8). 37 ANOVAs were performed, one for each *m/z* studied. If the between oil variance was significantly greater than the within oil variance (*F*-test at 95% probability), the *m/z* was identified as “potentially diagnostic”. For each potentially diagnostic *m/z* the classification percentage was determined from LSD values (significant at 99% level) on pairs of means of Frobenius norms. After investigating a range of thresholds (20%, 30%, 40%, 50%, 60%), 40% classification percentage was chosen as the optimum value. Thus any potentially diagnostic *m/z* with a classification percentage greater than 40% was identified as diagnostic, and then used for the PARAFAC2 analysis.

3.4.3. Implementation

All mathematical manipulation was performed on a WinTel personal computer (Pentium 4, 3.2 GHz) running Windows and Windows Office 2002 (Microsoft, Seattle, WA). PARAFAC2 version 1.011 was obtained from R. Bro at http://www.models.kvl.dk/source [16] and programmed in MATLAB, Version 7.1 (The MathWorks., Natick, MA). PCA using the Singular Value Decomposition (SVD) algorithm and other preprocessing methods (calculation of the first derivative and smoothing the EICs) was also implemented in MATLAB. ANOVA was performed using Analysis ToolPak Add-in in Microsoft Excel 2002. The PARAFAC2 routine in MATLAB was validated using the standard data set "chromatographic_fluo.zip" comprising shifted chromatographic data of thick juice samples from the beet sugar industry [15] available from http://www.models.kvl.dk/research/data/. The results agreed with the published data.

4. Results and discussion

4.1. Weathering effects

The weathering procedure applied in this work was dominated by evaporation and photo-oxidation processes. No effects due to biodegradation were anticipated as the samples were not in direct contact with microorganisms. The percentage loss of hydrocarbons (analytes) within the oil can be calculated relative to 17α(H)21β(H) hopane [10] which acts as an internal standard:

$$\%Loss = \frac{\left(A_F / H_F\right) - \left(A_w / H_w\right)}{\left(A_F / H_F\right)} \times 100 \qquad (2)$$

where A_F, H_F, A_w and H_w are the concentration (or peak area) of the analyte and 17α(H)21β(H) hopane in fresh and weathered samples respectively. To show the extent of weathering in this work the relative depletion of naphthalene (N), phenanthrene/anthracene (Ph/A, note that "/" does not represent a ratio here; Ph and A are characterized by the same *m/z* therefore are shown together), dibenzothiophene (DBT) and their one- to three-carbon side-chain substitutes were calculated for the example of a Kuwaiti crude oil. The result is shown in Figure 3.

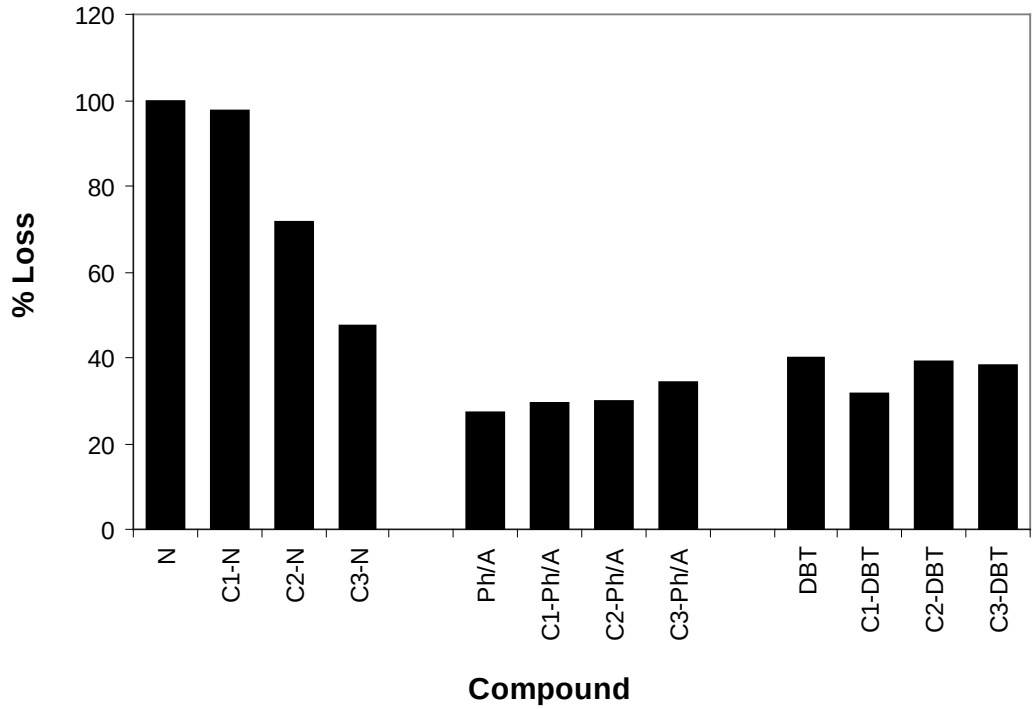

Figure 3. The relative loss (calculated by equation 2) of naphthalene (N), phenanthrene/anthracene (Ph/A), dibenzothiophene and their one- to three-carbon side-chain substitutes (C1 to C3) from the Kuwaiti crude oil after 15 days of weathering.

As indicated in Figure 3 these compounds exhibit depletion rates between 30 to 100% with maximum depletion rate in N group followed by DBT and Ph/A. The decline of weathering from N to C1-N indicates that the dominant weathering process to remove the N family is evaporation. On the other hand the reverse pattern of depletion in Ph/A and irregular pattern in the DBT family could have resulted from photo-oxidation. Contrary to evaporation, photo-oxidation increases with increasing carbon substitutes and number of rings [10]. A decline from Ph/A to C3-Ph/A and DBT to C3-DBT, which would be a result if evaporation were the major weathering process, is not seen in this experiment.

4.2. Exploratory data analysis

To speed up the PARAFAC2 algorithm and to avoid degeneracy [14,28,29] the elution time mode was decreased from 9670 to 100 data points by means of PCA as a data compression method before applying PARAFAC2 models [14]. Other methods such as those based on Fourier or wavelet transforms could also be used for the purpose of compressing the data in this step; however PCA was utilized as this method is available in almost all chemometrics software and is easy to apply. The new array of size 37 × 100 × 68 was decomposed using PARAFAC2 with one to eight factors (components). Fit and prediction values of the models revealed that five factors were adequate to model the significant variation of data (see Data Compression section and Figure S2 of the Supplementary Information for details). Classification of oil groups was achieved through visualization of the sample scores in the scores plot of PARAFAC2 components. Two samples found in the environment (DL1, DL2) were used in this study. In addition each of the weathered oil samples can be considered as an example of an environmental sample. Separated oil groups are shown in the four scores plots in Figure 4. At first sight, eight classified groups, including T, L, DL2, CK, CA, CM and diesel oils of DR and D1 are detectable; however samples D2-D8, CU and DL1 were not

classified or only partially classified. Oil sample groups CM and CA do not show tight groupings in the scores plots, although they have been reasonably separated from the other groups.

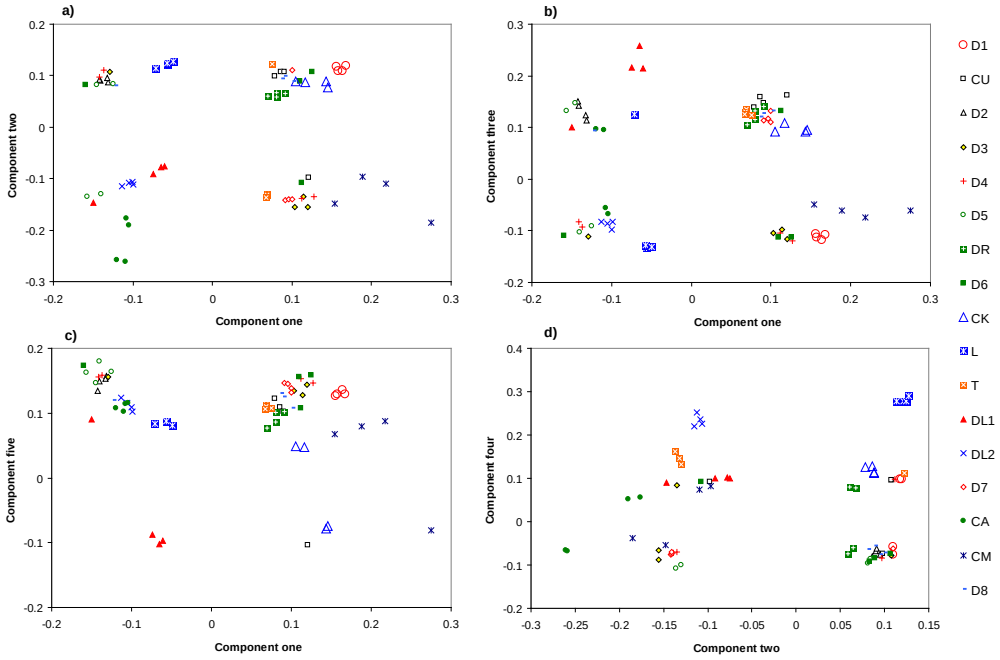

Figure 4. Scores plot of a PARAFAC2 model applied to the GC-MS data of 68 samples. a) component one-component two, b) component one-component three, c) component one-component five and d) component two-component four. See Table 1 for a description of the codes. CA, L, DL2, D1 and DR's classification is shown in (a) and grouped CM, T and CK in (b), (c) and (d) respectively.

Diesel oils and in particular those studied in this work have very similar GC-MS spectra and they can barely be distinguished by visual inspection. To show the degree of similarity between the diesel oils used in this study, the Total Ion Chromatograms (TIC) of D1, D4 and D5 are given in Figure S3 of the Supplementary Information. The rest of the diesel oils have similar spectra and have not been shown for brevity. D1 is the sample which was classified in the first PARAFAC run, D5 was partly classified and D4 was not classified at all. As displayed their TIC spectra are barely distinguishable by visual comparison. Therefore it is not surprising that samples D2-D8 are not grouped or partly grouped using an unsupervised PARAFAC2, however what is unexpected is that samples CU (archived as a crude oil) and DL1 (mixture of diesel and lubricating oils) were only partly grouped. By looking carefully into the 'Component 1-Component 5' scores plot (Figure 4c) it is apparent that except for one of the samples (which seems to be an outlier) in group DL1 the rest of the members of the group are classified far from the other groups as expected. Members of the CU group, on the other hand, are mixed with the other diesel groups in almost all the scores plots. It is suspected that CU was misclassified in the archive of samples at the NSWDEC. Further visual inspection of the GC-MS spectra of these two groups revealed that the non weathered oil sample of group DL1 exhibits quite different EICs in comparison to the weathered samples (2, 7 and 15 days weathered samples) in its group. The reason is not known but is under investigation. The CU group showed very similar composition (lacking high molecular weight biomarkers) and EIC patterns to those in the diesel oil samples. To assess if using a smaller data set could lead to an improved discrimination among oils, a new data set without samples T, L, DL2, CK, CA, CM, DR and D1 (i.e. groups that could be distinguished in the analysis of all samples) was analysed by PARAFAC2. The scores plot is depicted in Figure 5 showing two grouped samples of CU and DL1. The contention that group CU is a miss-

classified diesel is supported by its location close to other diesel samples. It is no surprise that no improvement in the discrimination of very similar local diesel oils D2-D8 is achieved.

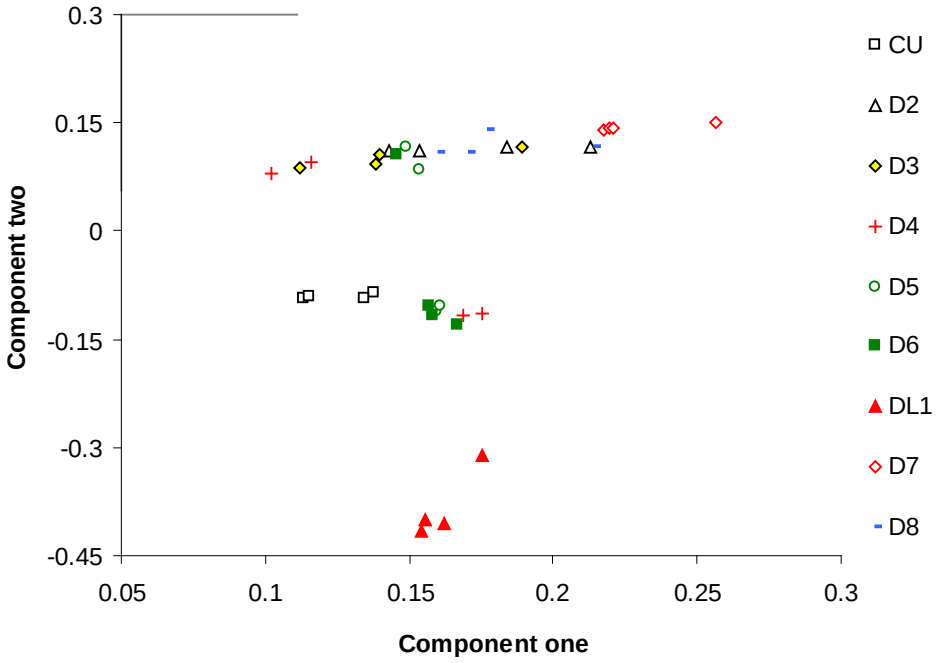

Figure 5. Scores plot of a PARAFAC2 model including CU, DL1 and D2 – D8 sample groups only.

4.3. Supervised pattern recognition with variable selection by ANOVA

Usually diesel oils exhibit similar EICs and their classification remains a challenge [11,30,31]. To improve the discrimination among D2-D8 diesel oils, the ANOVA variable selection procedure described above was used. 28 out of 37 *m/z* channels turned out to be potentially diagnostic. Applying LSD (at 99% probability level) to these *m/z* channels then returned eight *m/z* channels which displayed more than 40% classification percentage. These eight *m/z* characterize the following compounds: C2-naphthalenes (*m/z* 156), C4-benzo[b]thiophenes (190), C4-naphthalenes/dibenzothiophene (184), norhopanes (177), C3-benzo[b]thiophenes (176), C3-fluorenes (208), fluoranthene/pyrene (202) and C2-dibenzothiophenes (212). Naphthalenes, dibenzothiophenes and norhopanes are already known for being diagnostic among crude and heavy oils. However the proposed variable selection method indicates that fluoranthene/pyrene, C3-,C4-benzo[b]thiophenes and C3-fluorenes which have not been of much use, are important hydrocarbons that can be used in identification of very similar diesel oils.

Due to the high similarity between the diesel oil samples, no single *m/z* was found to differentiate between all the seven diesel oil groups and in the best situation, using C2-naphthalene (*m/z* 156), 14 out of 21 (66.7% classification percentage) comparisons turned out to be significant at 99% probability level. (A high degree of confidence was considered appropriate for this test, although testing at 95% gave the same result). The 37 *m/z* with the compound(s) they characterize and the classification percentage they return amongst the seven diesel oil groups of D2-D8 are given in Table S2 of the Supplementary Information. These results show that despite the high similarity between diesel oils there is a set of diagnostic fingerprints that can be used in the classification and identification of oil spills that have undergone a short term of weathering (up to 15 days).

Note that the aim of using LSD and setting the above mentioned criterion was to select those variables (*m/z*) which can provide the maximum separation among the seven diesel oil samples; that is, to present the highest number of separated classes when all the diesel oils are included in the analysis. If the problem were to distinguish between specified oils (or groups of oils) then the LSD can be targeted to the differences between the means of the specified oils. For example if oil D2 had to be discriminated from the rest (D3-D8) then there would be no need to calculate all pairwise LSDs, just those between D2 and each other oil group.

PARAFAC2 with 1 to 6 components was applied to a new data set of size 8 × 9669 × 28 embracing the first derivative and smoothed EICs of the above mentioned eight *m/z* for the 28 diesel oil samples (7 groups × 4 samples). Four components were shown to be adequate to model the systematic variation in the data. Components one, three and four were best for classification, while component two only explained overall variance. The classified oils are shown in a scores plot in Figure 6. As can be seen the oil groups of D6 and D7 were successfully classified. D4 and D5 classes show two rather spread groups which are separated from each other and other diesel oils except for D5. D2 and D8 are not totally separated, but they are not totally mixed groups either. After variable selection the informative sub-set of variables (EICs of eight diagnostic *m/z*), which exhibit greater differences between groups, are not overwhelmed by EICs representing almost similar patterns, leading to a better discrimination among difficult cases.

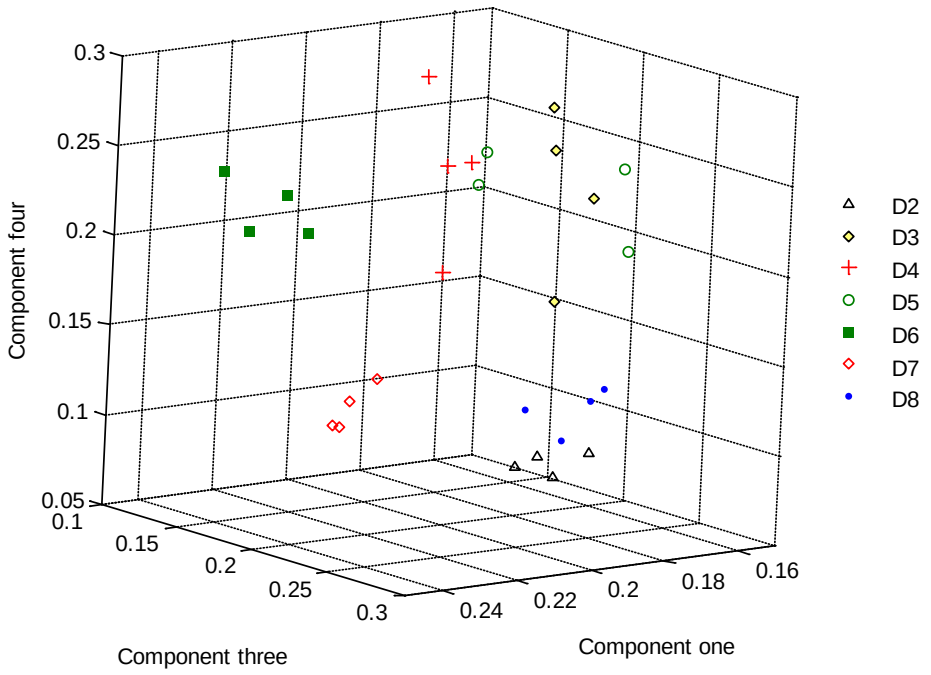

Figure 6. Scores plot of the PARAFAC2 model for D2 – D8 (see Table 1) with variable selection that includes only *m/z* 156, 190, 184, 177, 176, 208, 202 and 212.

5. Conclusion

This study represents the first application of PARAFAC2 for exploratory and supervised GC-MS analysis of different types of petroleum oils. The method can form the basis of an objective identification of the source of oil spills. The weathering experiment performed here was set up in a way to be dominated by photo-oxidation and evaporation which mimics the short term fate of an oil spill in a hot sunny weather (summer in Sydney). The Environmental Protection Authorities (EPAs) regularly monitor places with a risk of spillage of oil, and thus

15 days is a reasonable time to find and collect a spilled oil. However the chemometrics method described here could be used for more weathered samples providing that the EIC data of the highly altered and unstable chemical compounds are excluded from the analysis. The proposed PARAFAC2 method has been demonstrated to distinguish between some different short term weathered oil samples. The lack of a validation set does restrict the conclusions as to the general applicability of the method. However we have no evidence that the data used here were in any way peculiar and un-representative of the kinds of oils that would be found in the field. Due to the high similarity between diesel oils, a novel use of ANOVA and LSD was proposed as a variable selection tool to exclude non diagnostic ions from the analysis. Supervised pattern recognition then was shown to improve the classification power of the method. The proposed variable selection method revealed that less used hydrocarbons such as fluoranthene/pyrene, C3-,C4-benzo[b]thiophenes and C3-fluorenes are diagnostic among very similar diesel oils and therefore can be used in identification of their sources.

6. Acknowledgements

The authors thank Mr. Stephen Fuller from the NSW Department of Environment and Conservation (now Department of Environment and Climate Change) for the analysis of oil samples and his instructive comments. DE thanks the Australian Government for an International Postgraduate Research Scholarship.

7. References

[1] Z.D. Wang, M. Fingas, D.S. Page, Journal of Chromatography A 843 (1999) 369.

[2] ASTM, *D 5739-00 Standard Practice for Oil Spill Source Identification by Gas Chromatography and Positive Ion Electron Impact Low Resolution Mass Spectrometry*, American Society for Testing and Materials, Philadelphia 2000.

[3] J.H. Christensen, A.B. Hansen, G. Tomasi, J. Mortensen, O. Andersen, Environmental Science & Technology 38 (2004) 2912.

[4] P.S. Daling, L.-G. Faksness, A.B. Hansen, S.A. Stout, Environmental Forensics 3 (2002) 263.

[5] L.-G. Faksness, Weiss, H. M. and Daling, P. S., in SINTEF, Nordisk Innovations Center, Trondheim, 2002.

[6] J.M. Bayona, J. Albaiges, A.M. Solanas, R. Pares, P. Garrigues, M. Ewald, International Journal of Environmental Analytical Chemistry 23 (1986) 289.

[7] J.T. Andersson, Chemosphere 27 (1993) 2097.

[8] F. Jacquot, M. Guiliano, P. Doumenq, D. Munoz, G. Mille, Chemosphere 33 (1996) 671.

[9] B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, H.T. Mayfield, Analytica Chimica Acta 437 (2001) 233.

[10] R.C. Prince, R.M. Garrett, R.E. Bare, M.J. Grossman, T. Townsend, J.M. Suflita, K. Lee, E.H. Owens, G.A. Sergy, J.F. Braddock, J.E. Lindstrom, R.R. Lessard, Spill Science & Technology Bulletin 8 (2003) 145.

[11] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Environmental Forensics 7 (2006) 77.

[12] I.T. Jolliffe, Principal Component Analysis, Springer Verlag, New York, 2002.

[13] R. Bro, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149.

[14] R. Bro, Royal Veterinary and Agricultural University, 1998.

[15] R. Bro, C.A. Andersson, H.A.L. Kiers, Journal of Chemometrics 13 (1999) 295.

[16] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, Journal of Chemometrics 13 (1999) 275.

[17] R. Danielsson, D. Backstrom, S. Ullsten, Chemometrics and Intelligent Laboratory Systems 84 (2006) 33.

[18] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Journal of Chromatography A 805 (1998) 17.

[19] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, Journal of Chromatography A 961 (2002) 237.

[20] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, Journal of Chemometrics 17 (2003) 573.

[21] P. Wang, H. Tang, M.P. Fitzgibbon, M. McIntosh, M. Coram, H. Zhang, E. Yi, R. Aebersold, Biostatistics 8 (2007) 357.

[22] R.G. Brereton, Chemometrics Data Analysis for the Laboratory and Chemical Plant, Wiley, Chicester, 2003.

[23] D.B. Hibbert, J.J. Gooding, Data Analysis for Chemistry An Introductory Guide for Students and Laboratory Scientists, Oxford University Press, New York, 2005.

[24] K.J. Johnson, R.E. Synovec, Chemometrics and Intelligent Laboratory Systems 60 (2002) 225.

[25] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Analytical Chemistry 78 (2006) 5068.

[26] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Journal of Chromatography, A 1096 (2005) 101.

[27] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, Journal of Pharmaceutical and Biomedical Analysis 21 (1999) 115.

[28] A. Stegeman, Psychometrika 71 (2006) 483.

[29] B.J.H. Zijlstra, H.A.L. Kiers, Journal of Chemometrics 16 (2002) 596.

[30] Z. Wang, C. Yang, M. Fingas, B. Hollebone, X. Peng, A.B. Hansen, J.H. Christensen, Environmental Science and Technology 39 (2005) 8700.

[31] Z. Wang, C. Yang, B. Hollebone, M. Fingas, Environmental Science & Technology 40 (2006) 5636.