Journal of Chromatography A, 1235 (2012) 149–158

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Source identification of petroleum hydrocarbons in soil and sediments from Iguaçu River Watershed, Paraná, Brazil using the CHEMSIC method (CHEMometric analysis of Selected Ion Chromatograms)

Fabiana D.C. Gallotta[a,*], Jan H. Christensen[b,1]

[a] Environmental Assessment & Monitoring Department, Petrobras Research Center (CENPES), Petroleo Brasileiro S.A. – PETROBRAS, Av. Horácio de Macedo, 950 – Cidade Universitária, Ilha do Fundão ZIP: 21941-915 Rio de Janeiro, RJ, Brazil
[b] Department of Basic Sciences and Environment, The Faculty of Life Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 9 December 2011
Received in revised form 14 February 2012
Accepted 16 February 2012
Available online 23 February 2012

Keywords:
Polycyclic aromatic compounds (PACs)
Biomarkers
Petroleum hydrocarbons
Source identification
Chemometrics
Iguaçu River

ABSTRACT

A chemometric method based on principal component analysis (PCA) of pre-processed and combined sections of selected ion chromatograms (SICs) is used to characterise the hydrocarbon profiles in soil and sediment from Araucária, Guajuvira, General Lúcio and Balsa Nova Municipalities (Iguaçu River Watershed, Paraná, Brazil) and to indicate the main sources of hydrocarbon pollution. The study includes 38 SICs of polycyclic aromatic compounds (PACs) and four of petroleum biomarkers in two separate analyses. The most contaminated samples are inside the Presidente Getúlio Vargas Refinery area. These samples represent a petrogenic pattern and different weathering degrees. Samples from outside the refinery area are either less or not contaminated, or contain mixtures of diagenetic, pyrogenic and petrogenic inputs where different proportions predominate. The locations farthest away from industrial activity (Balsa Nova) contains the lowest levels of PAC contamination. There are no evidences to conclude positive matches between the samples from outside the refinery area and the Cusiana spilled oil.

1. Introduction

Hydrocarbons present in the environment consist of complex mixtures of compounds derived from multiple sources. The main contribution lies on fossil fuel inputs due to the rate and spatial scale by which petroleum has been used as an energy source and chemical feedstock. Petroleum hydrocarbons can be released into the environment through a number of pathways such as oil spills and natural seeps (petrogenic sources), and through incomplete combustion of fossil fuels (pyrogenic sources). Other contributions are hydrocarbons originating from biomass combustion; biosynthesis and early diagenetic transformation of non-hydrocarbon natural products to hydrocarbons [1–4].

Environmental assessment and monitoring programmes commonly focus on compliance-driven measurements, i.e. determining the concentration of concern hydrocarbon compounds. However, to assess correctly the significance of these concentration levels and to formulate adequate pollutant control strategies it is necessary to identify the source of the contamination.

A large number of techniques have been used for oil hydrocarbon fingerprinting and source identification [4–8]. Spectroscopic method are still used for screening in some oil spill investigations, but capillary gas chromatography (GC) clearly has proven most effective alternative [7]. Out of the GC methods, the use of diagnostic ratios of polycyclic aromatic compounds (PAC) and petroleum biomarkers are now widely accepted as an efficient technique for source identification [9,10]. The major drawbacks are the limited number of ratios assessed (e.g. 25 normative ratios in the CEN Guideline Method) [9] and the time-consuming and sometimes subjective task of integration (e.g. non-PACs incorrectly identified as isomers within an alkylated PAC group).

The combination of a large number of samples and many relevant compounds cited in the literature with the potential to distinguish sources brings out multivariate data analysis methods as a natural choice for data analysis. These methods allow for simultaneous analysis of many diagnostic ratios or normalised concentrations [6,11–13]. In this respect, the CHEMSIC method (CHEMometric analysis of sections of Selected Ion Chromatograms) suggested by Christensen et al. [14–16] but named for the first time in this study is a step forward, as combined sections of

* Corresponding author at: Environmental Assessment & Monitoring Department, PETROBRAS/CENPES/PDEDS/AMA, Av. Horácio de Macedo, 950 – Cidade Universitária, Ilha do Fundão ZIP: 21941-915 Rio de Janeiro, RJ, Brazil.
Tel.: +55 21 21626052; fax: +55 21 21624975.
E-mail addresses: fabianagallotta@petrobras.com.br (F.D.C. Gallotta), jch@life.ku.dk (J.H. Christensen).
[1] Tel.: +45 35332456.

0021-9673/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2012.02.041

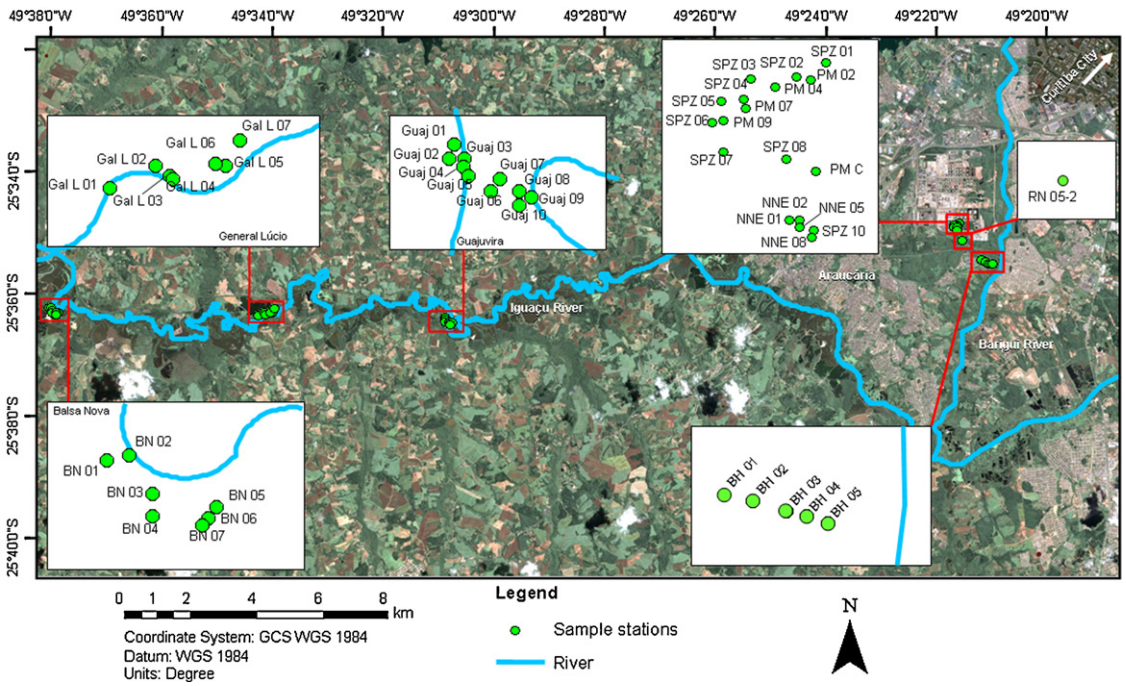

Fig. 1. Map of the study area. The sample labels abbreviations stand for sample type and location; inside the refinery area: PM (composed soil collected around a monitoring well); SPZ (surface soil); NNE (surface soil collected to the north of the new road); RN (surface soil collected nearby Saldanha Rivulet water-level scale); BH (composed soil collected around a monitoring well in a swampy area); and outside the refinery area: Guaj (surface soil in Guajuvira); Gal L (surface soil in General Lúcio) and BN (surface soil in Balsa Nova). These abbreviations have been used throughout in the environmental monitoring studies since 2000.

pre-processed selected ion chromatograms (SICs) are analysed directly by principal component analysis (PCA), without any prior peak integration or peak quantification. Therefore, unknown compounds or unique chemical features retained can lead to the recognition of new indicator compounds for source identification. Moreover, the method allows for the separation of the contributions from coeluting compounds and most likely increase the probability for separating samples with similar hydrocarbon composition. CHEMSIC method also provides significant advantages over chemometric methods based on total ion chromatograms (TIC). While CHEMSIC focuses on pre-selected groups of diagnostic compounds, TIC would mainly describe differences in the compounds present in higher concentrations (e.g. n-alkanes). Although the diagnostic compounds could be extracted from TIC, the lower sensitivity and especially the lower number of data points per peak would be major drawbacks to the pre-processing and PCA results.

In this study, soil and sediment samples from Araucária, Guajuvira, General Lúcio and Balsa Nova Municipalities (Iguaçu River Watershed, Paraná, Brazil) were analysed using the CHEMSIC method. The objectives were to characterise the hydrocarbon profiles for the study area and indicate the main sources of hydrocarbon pollution. Previously the CHEMSIC method has only been applied up to nine combined section [16]. Here, CHEMSIC was extended to include 38 SICs of PACs and four of petroleum biomarkers in two separate analyses. Moreover, different normalisation approaches were tested and discussed, focusing the chemometric analysis in different aspects of the chemical compositional variation within the data set. These novelties not only provide far more information on differences in the chemical composition of samples, but also on source identification and weathering behaviour of petroleum compounds. Finally, this study shows the method application and efficiency to a complex data set which includes samples at natural background level, samples with mixed contributions in different relative concentrations and crude oil samples.

2. Materials and methods

2.1. Sampling

Soil and sediment samples were collected in July 2009 and March 2010, using augers and cores from the refinery area and from points nearby the margins of Iguaçu River (Guajuvira, General Lúcio and Balsa Nova Municipalities). The samples were transferred into glass jars with Teflon caps, in which they were homogenised and maintained at −20 °C until preparation. Sampling location and labelling abbreviation details are shown in Fig. 1.

Araucária Municipality is located in a highly industrialised and urbanised region, in the vicinity of Curitiba City, where rivers Iguaçu and Barigüi receive intense chronic anthropogenic pollution. In the segment after Curitiba City, Iguaçu River is the second most polluted river in Brazil [17]. Moreover, on July 16th, 2000, approximately 4000 m^3 of Cusiana crude oil were released in the area, due to a pipeline rupture at the Presidente Getúlio Vargas Refinery, reaching the Saldanha Rivulet, the Barigüi River and subsequently the Iguaçu River [18].

Cusiana crude oil samples, two from the refinery tanks and one from the pipeline rupture point spill, were included for comparison with the environmental samples. These oil samples were collected in 2000 and have been stored into glass jars with Teflon caps and kept locked up in the dark at a maximum temperature of 4 °C. They are the most representative available samples of a non-weathered source of the spilled oil.

One oily water sample from an old monitoring well close to the pipeline rupture point was withdrawn through a bailer, transferred directly into a 40 mL glass jar with Teflon cap and maintained at 4 °C until preparation. The sample was expected to have higher concentrations of PAC and biomarkers, as oil vestiges were visible. It was included in the study to test the robustness of CHEMSIC method against subsequent dilutions.

2.2. Reagents and chemicals

Dichloromethane (Tedia, OH, USA) and *n*-hexane (Tedia, OH, USA) were all of pro analysis grade. Anhydrous sodium sulphate (Pro Analysis, Vetec, Rio de Janeiro, RJ, Brazil) was purified by heating at 400 °C for 4 h and allowed to cool in the desiccator. An aromatic surrogate standard mixture, containing naphthalene-d8, acenaphthene-d10, phenanthrene-d10, p-terphenyl-d14, chrysene-d12, perylene-d12 (AccuStandard, New Hawen, CT, USA), was added prior to the extraction. The instrument quality control mixture included DFTPP (decafluorotriphenylphosphine), 4,4′-DDT, pentachlorophenol, and benzidine (50 ng μL^{-1}, Supelco, Bellefonte, PA, USA).

2.3. Sample preparation

Aliquots of homogenised soil and sediment samples were mixed with anhydrous sodium sulphate and spiked with 50 μL of a 10 $\mu g\,mL^{-1}$ solution of the surrogate standard mixture. The amounts of material used for each extraction, and sodium sulphate for drying were selected according to the previous knowledge about contamination levels and the water percentage, respectively [19,20]. For samples with low level of total petroleum hydrocarbon (TPH) concentration (TPH < 3×10^2 mg kg^{-1}), 10 g of wet sample was homogenised with 30 g of sodium sulphate. For contaminated samples (TPH 3×10^2–1×10^4 mg kg^{-1}) the amount of material was reduced to 1 g of wet sample and 5 g of sodium sulphate.

Extracts of the solid samples were obtained according to EPA Method 3545A. Samples were extracted by pressurised liquid extraction using an ASE 300 instrument (Dionex, Sunnyvale, CA, USA). The method conditions were: extraction cells size 66 mL, extraction solvent dichloromethane, oven temperature 50 °C, pressure 1500 psi, pre-heating time 5 min, static time 5 min, flush volume of 10%, purge time 60 sec, and 3 static extraction cycles. After extraction, the solvent volume was reduced to 1 mL using a TurboVap 500 (Calliper Life Science, Hopkinton, MA, USA), 30 mL of *n*-hexane was added and the extract concentrated to 1 mL.

SPE cartridges packed with 1.5 g silica and 6 g cyanopropyl (Interchim, Montluçon, France) was used for cleanup. After preconditioning of the columns with 6 mL of *n*-hexane, the extracts were applied to the top of the column, and petroleum hydrocarbons eluted with 10 mL of dichloromethane in *n*-hexane (1:9, v:v). The eluent was then concentrated to 1 mL and maintained at −20 °C in amber glass vials until analysis.

The oily water sample (around 40 mL) was spiked with 50 μL of the surrogate standard mixture and liquid–liquid extracted 3 times with 5 mL of *n*-hexane. The combined extract was filtered through a funnel filled with sodium sulphate, concentrated to 1 mL and treated with the same SPE procedure as sediment samples. Cusiana oil samples (1 mL of 2000 mg L^{-1} oil solution in *n*-hexane) were spiked with the surrogate standard mixture and treated by the same SPE procedure as the sediment samples.

2.4. GC–MS analysis

The samples were analysed using an Agilent 6890N/5975 GC–MS operating in electron ionisation mode. The GC was equipped with a 60 m ZB-5 (0.25 mm I.D., 0.25 μm film thickness) capillary column. Helium was used as carrier gas with a flow rate of 1.1 mL s^{-1}. Aliquots of 1 μL were injected in pulsed splitless mode with injection temperature of 315 °C. The column temperature programme was as follows: Initial temperature 40 °C held for 2 min, 25 °C min^{-1} to 100 °C then followed by an increase of 5 °C min^{-1} to 315 °C (held for 13.4 min). The transfer line, ion source and quadrupole temperatures were 315 °C, 230 °C and 150 °C, respectively. A total of 55 mass-to-charge ratios (*m*/*z*'s) divided into 12 groups were acquired in SIM mode (cf. Table 1). The dwell time for each *m*/*z* was 25 ms with 2.81 scans s^{-1}. The number of monitored ions (13 *m*/*z*'s) was consistent between groups to avoid differences in the scanning frequency.

2.5. Quality control

The samples were divided into 6 batches. In the analytical sequence, dichloromethane, an oil sample (1:1 mixture of heavy fuel oil from the Baltic Carrier and North Sea crude oil from the Brent oil field [6]) and the instrument quality control mixture were analysed between batches. These test solutions were used for quality control by daily monitoring for cross-contamination; changes in peak shapes, chromatographic resolution and sensitivity; and to verify tuning, injection port inertness and GC column performance, respectively.

Besides the regular quality control of GC–MS methods, when using chemometric data analysis of sections of chromatograms, additional validation samples are highly recommended, i.e. analytical replicates spread into the batches. These samples are used to ensure that the data processing is able to remove the variation unrelated to the chemical composition. Five sample extracts were included randomly in each batch: Cusiana oil sample, two samples with a weathered oil fingerprint, one sample with a high carbon preference index (CPI) value and absence of unresolved complex mixture (UCM), and one sample with both a high CPI value and UCM. The choice was based on previous GC-FID analysis [19,20].

Additionally, a 'reference' sample extract consisting of equal amounts of eight extracts was analysed every ten runs. In this case study, previous GC-FID analyses were available, facilitating the selection of sample extracts representative of the whole sample set.

Furthermore, sample dilutions (1:10, 1:20, 1:40 and 1:100) of the most concentrated sample (oily water) were prepared and injected in the last batch. These dilutions were used to indicate how reproducible the CHEMSIC method is to cluster/identify samples with the same chemical composition but with different TPH concentrations.

2.6. Data

The data set consisted of retention time windows of 55 SICs per sample (cf. Table 1), including the deuterated standards.

A total of 127 samples were analysed and split into a 'training set' of 66 sample extracts and four 'validation sets'. The training set included samples collected from the study area (cf. Fig. 1) and samples of Cusiana oil. The four validation sets were: eight sampling duplicates ('Duplicates' in PCA plots), four dilutions of a concentrated sample ('Dilutions' in PCA plots), 19 replicate analyses of the reference sample extract ('References' in PCA plots), and 30 analytical replicates (six injections of five selected sample extracts, 'Replicates' in PCA plots).

2.7. Data processing and analysis

The data consisting of 55 GC–MS/SIM chromatograms for each sample were exported to the AIA file format using the commercial software ChemStation (Agilent Technologies). NetCDF was used to retrieve relevant data (e.g. signal intensities, sample names, sample descriptions) in the MATLAB 7.10.0 (R2010a) programming environment, in which the data were pre-processed and analysed. The algorithms for import of CDF-files, correlation optimised warping (COW) and PCA were downloaded from www.models.life.ku.dk.

The chromatograms comprising between 344 and 8710 data points, were reduced before data processing, by visual inspection, eliminating parts with no relevant information. The parts removed included sections of the chromatogram with low

Table 1
List of compounds, SICs and corresponding groups of GC/MS-SIM.

Compounds	SIC	Group(s)	Compounds	SIC	Group(s)
n-Alkyl cyclo hexanes	83	I to XII	C4-Decalins C2-fluorenes[d]	194	I + II + III + IV + VI + VII
Alkanes	85	I to XII	C2-Dibenzofurans[d]	196	V + VI + VII
Alkyl toluenes	105	I to XII	C1-Dibenzothiophenes[c,d]	198	VI + VII
Sesquiterpanes	123	I to VI	C0-Fluoranthene[c,d] C0-Pyrene[c,d]	202	VII + VIII + IX
Naphthalene[d]	128	I	C2-Phenanthrenes/anthracenes[d]	206	VII + VIII + IX
Benzo(b)thiophene[d]	134	I	C3-Fluorenes[d]	208	VI + VII
d8-Naphthalene[b]	136	I	C2-Dibenzothiophenes[d] d10-Fluoranthene[a] d10-Pyrene[a]	212	VII + VIII
C0-Decalin	138	I	C1-Fluoranthenes/pyrenes[c,d]	216	VIII + IX
C1-Naphthalenes[d]	142	II	Steranes	217	VIII to XII
C1-Benzo(b)thiophenes[d]	148	I + II	Steranes	218	X + XI + XII
C1-Decalins Acenaphthylene[d]	152	I + II + III	C3-Phenanthrenes/anthracenes[d]	220	VII + VIII + IX
Acenaphthene[d]	154	II + III + IV	C3-Dibenzothiophenes[d]	226	VII + VIII + IX
C2-Naphthalenes[d]	156	III	C0-Benzo(a)anthracene[c,d] C0-Chrysene[c,d]	228	X
d8-Acenaphthylene[a]	160	III	C2-Fluoranthenes/pyrenes[d]	230	IX + X
C2-Benzo(b)thiophenes[d]	162	II + III	Triaromatic steranes	231	X + XI + XII
d10-Acenaphthene[b]	164	III + IV	C4-Phenanthrenes/anthracenes[d] Retene[d] C0-Benzonaphtothiophene[d]	234	VIII + IX + X
C2-Decalins C0-Fluorene[d]	166	I + II + V	C4-Dibenzothiophenes d12-Benzo(a)anthracene[a] d12-Chrysene[b]	240	VIII + IX + X
C0-Dibenzofuran[d]	168	II + III + IV	C1-Chrysenes[c,d]	242	X + XI
C3-Naphthalenes[d]	170	IV + V	d14-p-Terphenyl[b]	244	VIII
C3-Benzo(b)thiophenes[d] d10-Fluorene[a]	176	IV + V	C1-Benzonaphtothiophenes[d]	248	X + XI
C0-Phenanthrene [c,d] C0-Anthracene[c,d]	178	VI	5 Rings PAHs[c,d]	252	XI + XII
C3-Decalins C1-Fluorenes[d]	180	I + II + III + V	C2-Chrysenes[d]	256	XI
C1-Dibenzofurans[d]	182	IV + V + VI	d12-Benzo(k)fluoranthene[a] d12-Benzo(a)pyrene[a] d12-Perylene[b]	264	XI + XII
C4-Naphthalenes[d] C0-Dibenzothiophene[d]	184	IV + V + VI	C3-Chrysenes[d]	270	XI + XII
d10-Phenanthrene[b] d10-Anthracene[a]	188	VI	6 Rings PAHs[c,d]	276	XII
C4-Benzo(b)thiophenes[d]	190	IV + V	6 Rings PAHs[d]	278	XII
Tricyclic terpanes Hopanes	191	IX + X + XI + XII	d12-Indeno(1,2,3-cd)pyrene[a] d12-Benzo(g,h,i)perylene[a]	288	XII
C1-Phenanthrenes/anthracenes[c,d] d8-Dibenzothiophene[a]	192	V + VI + VII			

[a] Internal standards analysed but not added to the samples in this particular case study;
[b] internal standards used for normalisation **Scheme I** in the initial source identification based on a subset of SICs;
[c] SICs included for the initial source identification based on a subset of SICs;
[d] SICs included for source identification using relative fingerprints of 38 groups of PACs.

signal-to-noise ratio (S/N) and sections where target compounds were not expected. The CHEMSIC procedure described by Christensen et al. [14–16] was utilised in this work aiming at taking away variation that is unrelated to the chemical composition. The preprocessing consists of baseline removal, retention time alignment and data normalisation.

Briefly, the baseline was removed by calculating the first derivatives of the chromatographic data (point-by-point subtraction). The retention time alignment was performed in two steps: (i) applying rigid shifts (i.e. without compression or expansion) on the chromatograms, and (ii) employing the COW algorithm [21,22]. The COW algorithm aligns a sample chromatogram towards a target chromatogram by stretching or compressing sample segments along the retention time axis using linear interpolation. The optimal warping parameters (i.e. the length of the segments, in which the signals are divided, and 'slack parameter', how much it is allowed to change) were determined by the use of a grid search in the parameter space followed by a discrete simplex-search on maximum values for the 'warping effect' function [23].

The SICs for each m/z were always aligned separately to the SICs of a target sample. The target for the alignment was selected from the reference samples, using the one with the highest sum of correlation coefficients with the others. The reference samples were chosen in this study as they were prepared to be an average sample containing most peaks.

Although baseline removal and retention time alignment are essential steps to prepare the data, the methods involved are not important for the further interpretation of the model. However, data normalisation affects the interpretation of the results and focus of the subsequent data analysis. The aims of the normalisation step are to remove variations unrelated to the chemical information such as time related changes in sensitivity; and to focus the subsequent chemometric data analyses on different types of aspects (viz. compound concentrations, differences between groups of compounds (SICs) or differences in relative concentrations within SICs).

In this study, we therefore applied three normalisation schemes that focus on these three types of variation in data. In **Scheme I**, we normalise each SIC to an internal standard. The deuterium

labelled internal standard with most similar physicochemical properties was selected for each SIC (e.g. d10-phenanthrene for C1 and C2-phenanthrenes). Here variations related to sample preparation and instrumental analyses are reduced, but all other chemically relevant information is retained. This normalisation scheme will typically focus the analysis on variations in total hydrocarbon concentrations, followed by variations between and within groups of compounds (SICs) as the former will typically be the most pronounced variation found by PCA of (semi)quantitative data [15]. In **Scheme II**, one more layer of chemical information (i.e. total hydrocarbon concentrations) is removed. This is done by combining SICs and then normalising to constant Euclidean norm (i.e. corresponds to normalisation to the sum, if data were consisting of only positive values) [16]. This normalisation scheme focuses the analysis on variations between SICs followed by variations within SICs. Finally, in **Scheme III**, data are normalised to constant Euclidean norm within each SIC and then SICs are combined. Thus, the PCA will focus solely on chemical variations within each SIC such as differences in isomer PAC patterns and biomarker fingerprints.

The pre-processed signals for all the samples were stacked in an $I \times J$ matrix X, where I denotes the number of samples and J the length of the signals, and modelled by PCA. The PC model was calculated on the column-wise centred training set and the validation sets (viz. samples that are not present in the training set) were projected on it after centring to the mean of the training set. The model was further validated by visual inspection of the loadings and chemical interpretation of the scores and loadings.

3. Results

3.1. Initial source identification based on a subset of SICs

Out of the 55 *m*/*z*'s analysed, SICs of nine groups of PACs relevant for source identification (marked with 'c' in Table 1 and shown in Fig. 2a) were used for this initial study [16]. Additionally, SICs of selected internal standards (marked with 'b' in Table 1) were also pre-processed and used for normalisation in **Scheme I**. The baseline was removed by calculating the first derivative of the SICs. The retention time shifts in the data set were between 7 and 20 scan points, depending on the SIC. The rigid shift procedure took care of the main part of the constant shift within each SIC. For SICs with less than 200 scan points, the search for the optimal segment length was between 25 and 50 scan points (with 5 point increments) and for SICs with more than 200 scan points, the range was between 50 and 100 scan points (with 10 point increments). The slack parameter grid was 1–3 with 1 point increment for all SICs. The maximum correction allowed was 5 scan points. The optimal segment lengths were between 26 and 89 scan points and the optima for the slack parameter were between 1 and 3 points.

The effect on data after the application of the normalisation schemes is shown in Fig. 2. In **Scheme I** (Fig. 2b–e), data exhibit quite different intensity ranges (maximum from 1.5 up to 250), which reflect the variations in absolute PAC concentrations. Fig. 2c shows the most concentrated sample. On the other hand, the intensities are in the same order of magnitude after normalisation to **Scheme II** (Fig. 2f–i) and **Scheme III** (Fig. 2j–m). In **Scheme II**, the increasing physical weathering from Fig. 2f–i is evident as a relative decrease of low-molecular-weight (low-MW) PACs (viz. *m*/*z*'s 178, 192 and 198 SICs) compared to high-molecular-weight (high-MW) PACs. Conversely, the use of normalisation **Scheme III** removes the relative difference between the concentrations of each group of PACs. It instead brings forward differences in the patterns of isomers within each SIC. Hence, the differences between fresh oil (Fig. 2j) and moderated weathered samples are clearly reduced by normalisation **Scheme III**. Although the differences in relative amounts of individual isomers within the alkylated families (e.g. *m*/*z*'s 192 and 198) are retained in all schemes (Fig. 2b–m), in **Scheme III** (Fig. 2j–m) these differences constitute the major variation.

3.1.1. Pollution levels and weathering

After applying normalisation **Scheme I** (normalisation to internal standards), the SICs were combined (e.g. Fig. 2b–e). Each one of the 66 samples now consisted of 1839 data points. A bend in explained variance was observed past the third PC for the training set. Furthermore, since the loadings above PC3 contain shift patterns (Fig. S1) in addition to chemical variation, it was concluded that the optimal number of PCs is three [15]. The three-component PCA model describes 98.6% of the variance.

The model provides information on relative contamination levels (relative to an average sample) included in the training set and information on PAC sources. The average sample is a slightly weathered crude oil, but with high levels of perylene (Fig. S2). The PC1 loading is similar to the average chromatogram, except for perylene which has a small negative loading coefficient and other 5- and 6-ring PACs which have lower positive PC1 loadings than the mean chromatogram (Fig. S2a). In this study, larger positive PC1 scores correspond to high concentrations of weathered crude oil, while samples with large negative PC1 scores are the least contaminated. Samples with PC1 scores around zero have an average contamination level for the sample set (Fig. 3). Samples with highest PC1 scores (e.g. BH01) were removed from the 'training set' to test outliers. The model was recalculated and did not show changes, therefore all samples were retained.

Although PCA cannot provide an accurate apportionment of the different sources of PAC contamination, it can be used to determine the main sources of PAC contamination. Firstly, as the PC2 loading coefficients are negative for (low-MW) PACs such as phenanthrene and 3- and 2-methylphenanthrene (Fig. S2b), it can be concluded that samples with large negative PC2 scores (Fig. 3) contain higher concentrations of relatively fresh mineral oil. Likewise, as the loading coefficients are close to zero for anthracene, C1-anthracenes, fluoranthene and C1-fluoranthenes, and positive for high-MW PACs such as pyrene, C1-pyrenes, chrysene, C1-chrysenes, benzo(e)pyrene and benzo(g,h,i)perylene (Fig. S2b), samples with large positive PC2 scores (Fig. 3) contain higher concentrations of a moderately weathered oil (or heavy fuel oil without cracking residues). These conclusions can be drawn as high relative concentrations of phenanthrene and C1-dibenzothiophenes; and of pyrene, C1-pyrenes, chrysene, C1-chrysenes, benzo(e)pyrene and benzo(g,h,i)perylene are indicators for low and high-MW petrogenic input, respectively [4]. It is also noteworthy that among the C1-phenathrenes, the 3- and 2-methyl isomers have negative PC2 loading coefficients, while the 9/4- and 1-methyl present positive coefficients. Although these coefficients are small, this pattern in the C1-phenathrenes gives some indication of biodegradation [14].

The PC3 loading coefficients are close to zero for all the compounds except for perylene, which has a large negative value (Fig. S2c). PC3 can therefore be interpreted as an indicator for in situ diagenesis. Samples with large negative PC3 scores (Fig. 3) therefore have a large diagenetic input.

Samples with positive PC1 scores and negative PC2 scores contain relatively fresh crude oil, e.g. PM 02, SPZ 08, NNE 08, BH 01 (marked with * in Fig. 3) when compared to the average sample. Conversely, samples with positive PC2 scores, e.g. BH03, BH04 (marked with § in Fig. 3), contain higher concentrations of weathered crude oil. It is noteworthy that the Cusiana oil samples (marked with † in Fig. 3) have PC1 scores close to zero and negative PC2 scores, which identify a typical fresh crude oil PAC pattern with average oil concentrations, as expected.

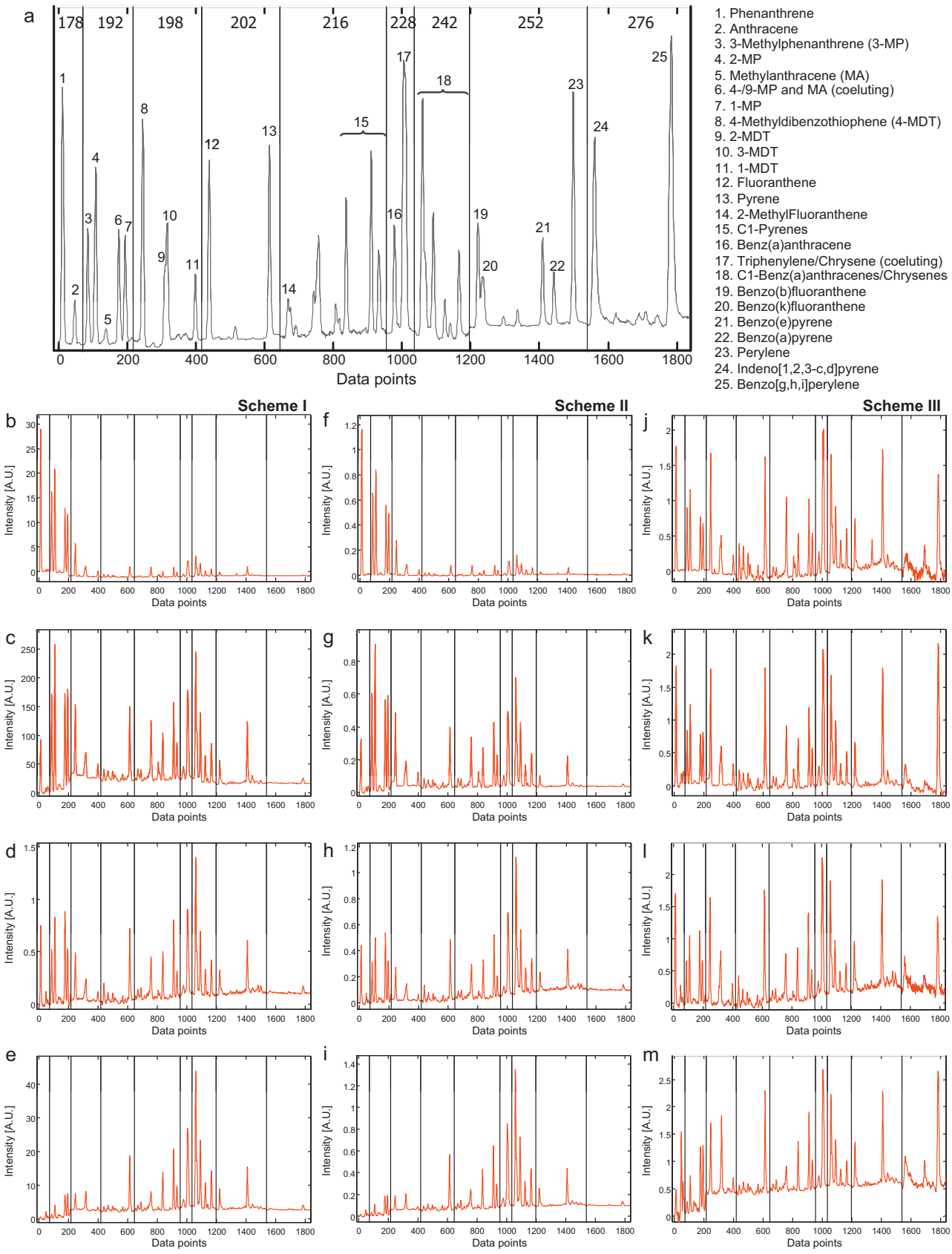

Fig. 2. Data Normalisation: (a) Peaks or peak clusters identification; **Scheme I** (normalisation to internal standards): (b) Cusiana oil, (c) BH 02, (d) BH 04, (e) BH 03; **Scheme II** (concatenation and normalisation to Euclidean norm): (f) Cusiana oil, (g) BH 02, (h) BH 04, (i) BH 03; **Scheme III** (normalisation to Euclidean norm within SICs): (j) Cusiana oil, (k) BH 02, (l) BH 04, (m) BH 03. Note that the SICs on the top of Fig. 2a are the same showed in the other figures and some were omitted to improve the readability.

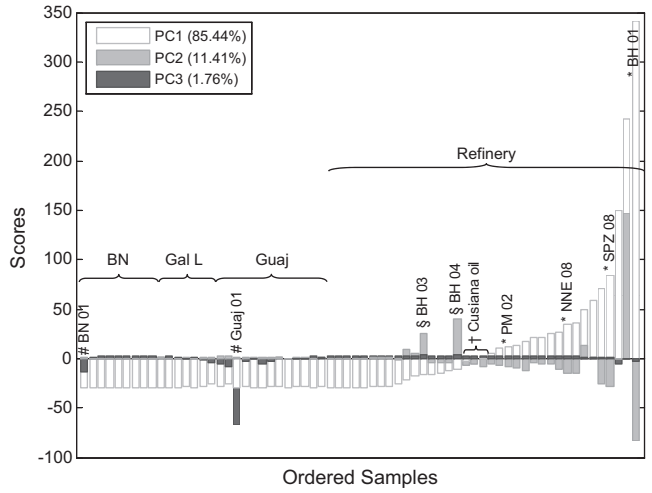

Fig. 3. Subset of 9 PACs data analysis: Scores for PC1, PC2 and PC3 for the principal component model of mean-centred data of normalisation **Scheme I** (normalisation to the internal standards). BN, Gal L, Guaj and Refinery denote samples from Balsa Nova, General Lúcio, Guajuvira and from inside the Refinery area, respectively. The symbols indicate samples: less (*) and more (§) weathered than the mean sample, with significant diagenetic input (#) and the Cusiana oil (†). Note that bars are not stacked.

The samples with negative PC1 scores and negative PC3 scores, e.g. BN 01 and Guaj 01 (marked with # in Fig. 3), are among the least contaminated and contain a large amount of perylene. It is, therefore, evident that these samples represent important diagenetic input, while PACs of petrogenic input are limited.

Samples from outside the refinery area are the least contaminated, i.e. samples from Guajuvira ('Guaj' in Fig. 3), General Lúcio ('Gal L' in Fig. 3) and Balsa Nova ('BN' in Fig. 3), while soil samples from inside the refinery area ('Refinery' in Fig. 3) are more contaminated than the mean sample.

3.1.2. Exclusion of samples with low contamination level

Principal component models on the entire data set (66 sample extracts × 1839 data points) using all three normalisation schemes revealed that a number of samples (BN 02, BN 03, BN 04, BN 05, BN 06, BN 07, SPZ 02, SPZ 07, BH 01 and BH 05) contain very low levels of all PACs. These samples were therefore removed to focus the analyses on the source of contaminants for the remaining samples. Fig. 4 shows the score plot of PC1 vs. PC2 for the principal component model on data normalised to Euclidean norm within each

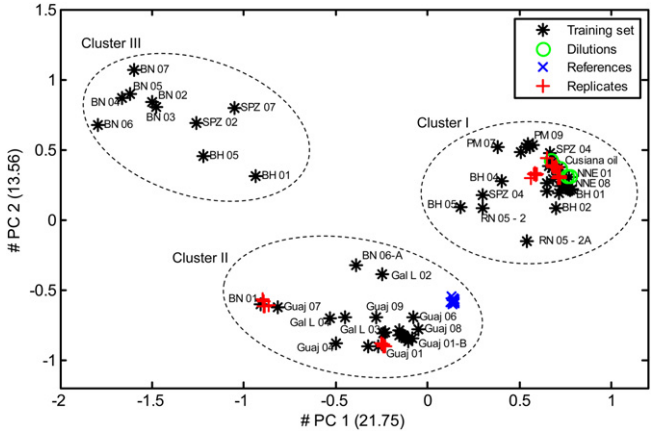

Fig. 4. Subset of 9 PACs data analysis: Score plot of PC1 vs. PC2 for the principal component model of normalisation **Scheme III** (normalisation to Euclidean norm within SICs). Some labels were omitted from the Score plot to improve the readability.

SIC (**Scheme III**). Chemical interpretation of the loading plots for this model (not shown) and of the SICs for individual samples in the three clusters (Fig. S3a–e) reveal that samples in cluster I are mainly of petrogenic origin. These samples are all from inside the refinery area. Conversely, samples in cluster II are all from outside the refinery area (Guajuvira, General Lúcio and Balsa Nova). These samples are predominated by PACs from a mixture of diagenetic (i.e. perylene, Fig. S3c: BN 01) and pyrogenic origin (i.e. fluoranthene, benzo(a)anthracene, high-MW non-alkylated PACs, Fig. S3d: Gal L 02). Finally, there were no PAC peaks present in the SICs of samples located within cluster III. A typical example of combined SICs for one of these samples are shown in Fig. S3e (BN 02). Samples located in cluster III were therefore excluded from the subsequent chemometric analyses, as they contain no information.

3.2. Source identification using relative fingerprints of 38 groups of PACs

A total of 38 SICs containing 2–6 ring PACs (marked with 'd' in Table 1) were included as variables in the model. After concatenation each one of the 56 samples consisted of 10,738 data points. The baseline removal was performed as described in Section 3.1. Normalisation **Scheme II** (concatenation and normalisation to Euclidean norm) was applied as last step of the pre-processing. A three-component model describing 88.3% of the variance in the training set was found to be optimal.

The score plot in Fig. 5a shows PC1 vs. PC2 for this model. PC1 distinguishes the petrogenic samples from the samples presenting mainly diagenetic and pyrogenic inputs, as the loading coefficients (Fig. 5b) are positive for compounds with a primarily petrogenic origin (i.e. naphthalenes, phenanthrenes, dibenzothiophenes and fluorenes) and especially large and positive for C0–C3 naphthalenes and C0–C2 phenanthrenes. Fig. 5d is a zoom of the PC1 loading coefficients of fluoranthene, pyrene, benzo(a)anthracene, chrysene, 5- and 6-ring PAC, where the typical diagenetic (i.e. perylene) and pyrogenic (i.e. fluoranthene, benzo(a)anthracene, indeno(1,2,3-cd)pyrene) compounds have negative coefficients. Perylene has the largest negative loading coefficient (large ratio to other 5-ring PACs), indicating a significant diagenetic background [24] for samples outside the refinery area, which have negative PC1 score values. Moreover, the ratios of fluoranthene to pyrene, benzo(a)anthracenes to chrysene and indeno(1,2,3-cd)pyrene to benzo(g,h,i)perylene (see Fig. 5d, where each pair of these peaks has approximately the same size) evidence pyrogenic input for the samples clustering in the bottom left of Fig. 5a as well [4]. This cluster includes all samples from outside the refinery area, i.e. samples from Guajuvira, General Lúcio and Balsa Nova. PC2 scores are close to zero or have small negative values, therefore having low influence in the interpretation of results.

Most of the samples from inside the refinery area are located on a trajectory from positive PC1 and negative PC2 scores such as the Cusiana oil (Fig. 5a, bottom right), to negative PC1 and large positive PC2, e.g. BH 04 (Fig. 5a, top left ellipse). PC2 loading coefficients (Fig. 5c) are negative for C0- to C2-naphthalenes and C0-phenanthrene, close to zero for C0- and C1-fluorenes and positive for C3- and C4- naphthalenes, C2- to C3-fluorenes and C1- to C4-phenanthrenes. The observed effects on the distribution among C0–C4 alkylated families indicate physical weathering, but could also be caused partially by biodegradation processes as an increase in alkylation level decreases the susceptibility to microbial attack [25]. Fig. 5e is a zoom of the PC2 loading coefficients for C0- to C3-naphthalenes, where the loss of the low-boiling-point compounds is evident with increasing PC2 scores as C1- and C2-naphthalenes have negative PC2 loading coefficients and C3-naphthalenes have positive ones. Hence, BH 04 and BH 03 are among the most weathered samples (Fig. 5a, top left ellipse), while SPZ 05, SPZ 08, PM 02,

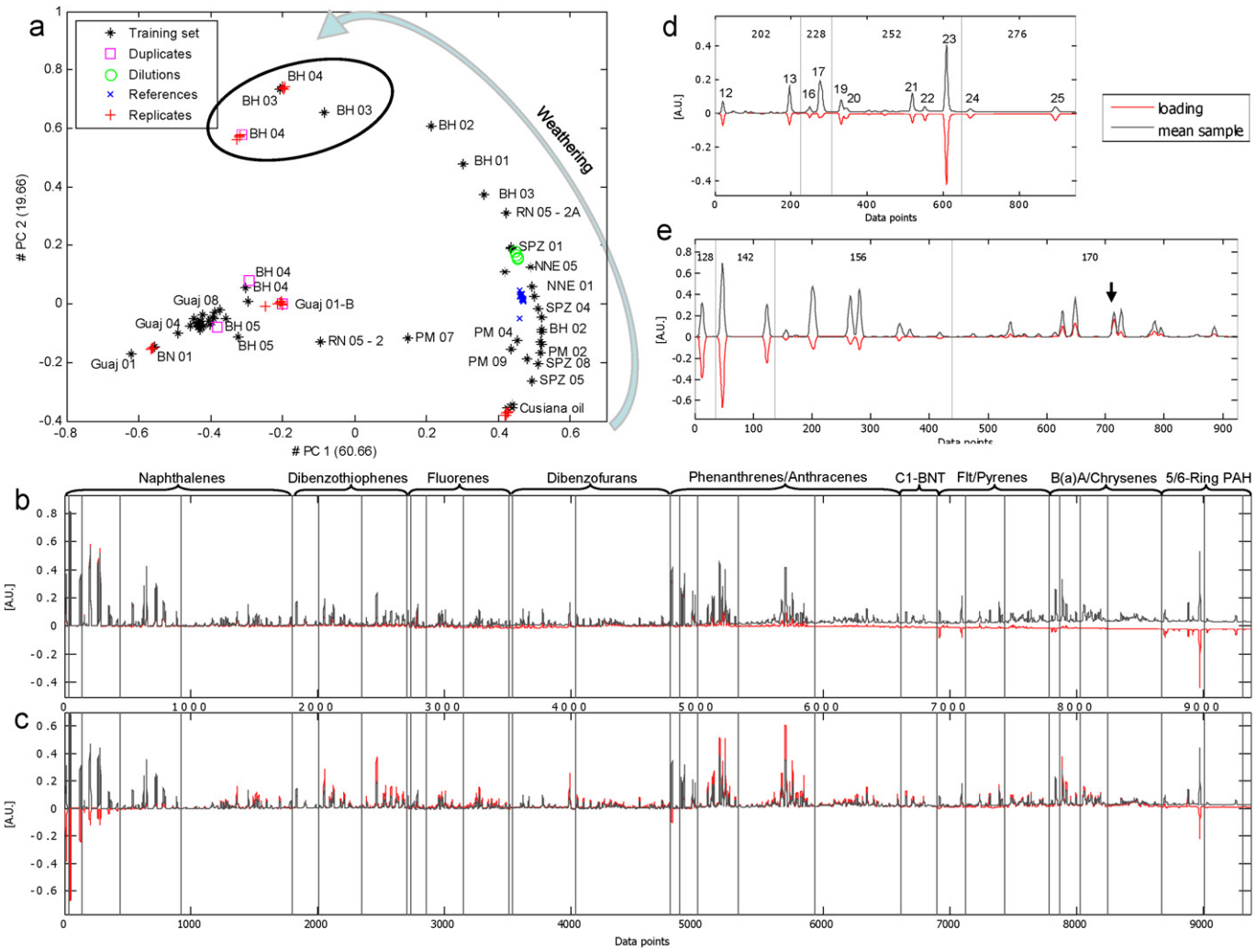

Fig. 5. 38 PACs data analysis: (a) Score plot of PC1 vs. PC2 for the principal component model of normalisation **Scheme II** (concatenation and normalisation to Euclidean norm), Loading plots: (b) PC1 and (c) PC2. The dotted lines are the mean chromatogram of the entire training set, while the solid lines are the loadings (A.U.: Arbitrary Units), (d) Zoom of PC1 loadings for pyrogenic compounds, (e) Zoom of PC2 loadings for naphthalenes (C4-naphthalenes were not included to improve the visualisation). Some labels were omitted from the Score plot (a) to improve the readability. The SICs corresponding to C0-C4-Benzothiophenes, acenaphthylene and acenaphthene were omitted from the Loading plots (b) and (c) to improve the visualisation. For these compounds the loading coefficients are close to zero.

PM 09 and PM 04 (Fig. 5a bottom right) are relatively fresh crude oil with similar PAC composition as the Cusiana oil. The arrow in Fig. 5a emphasizes the direction of increased weathering degree. All samples following this trend were collected close to the spill accident which provides additional evidence that these samples contain a weathered fraction of the Cusiana oil. Besides the general physical weathering trend, changes in isomeric variations within alkylated PACs families indicate that biodegradation has indeed occurred. Further investigation, including just the oil contaminated samples and selecting solely alkylated homologues as variables for the model, will be treated in details in a separated paper.

Perylene has the most prominent positive PC3 loading coefficient (Fig. S4), which gives some information on the diagenetic input. In addition, C2- and C3-naphthalenes, C0- and C1-fluorenes and C0- and C1-phenanthrenes have positive PC3 loading coefficients while more alkylated (C2–C4) isomers and 4-ring PACs have negative coefficients. This demonstrates that PC3 is a mixed component that mostly explains diagenetic input and intermediate weathered oils and that some samples contain significant levels of both diagenetic and petrogenic input. Therefore, although uncorrected retention time shifts and peak shape changes are not visible, a clear interpretation of the component is hindered by the mixture of contributions.

3.3. Biomarkers data analysis (4 SICs)

Terpane; regular and dia-sterane; and triaromatic sterane SICs (Table 1) were included as variables in the model. After baseline removal and alignment of 7654 data points per sample ('training set' of 56 samples × 7654 data points), normalisation **Scheme III** (Euclidean norm within each SIC) was applied. This scheme ensures focus on the relative composition within each of the biomarker groups. A three-component model describing 32.1% of the variance in the training set was found to be optimal as a clear bend in the explained variance of the validation set is observed and since the loadings above PC3 contain shift patterns. Fig. 6a shows the score plot of PC1 vs. PC2 for the PC model, where two clusters can be identified.

Samples from the refinery area cluster in the upper right corner of the score plot (Fig. 6a, Cluster I) with positive PC1 scores and around zero and positive PC2 scores. These samples have biomarker profiles with the highest similarity to the Cusiana oil (Fig. 6b). N.B. the Cusiana oil is located in the cluster of samples from the refinery area. The characteristics of these samples are among others that they have a high relative concentration of tricyclic terpanes, hopanes, particularly 17α(H),21β(H)-30-norhopane (H29) and 17α(H),21β(H)-hopane (H30), and in general a high relative

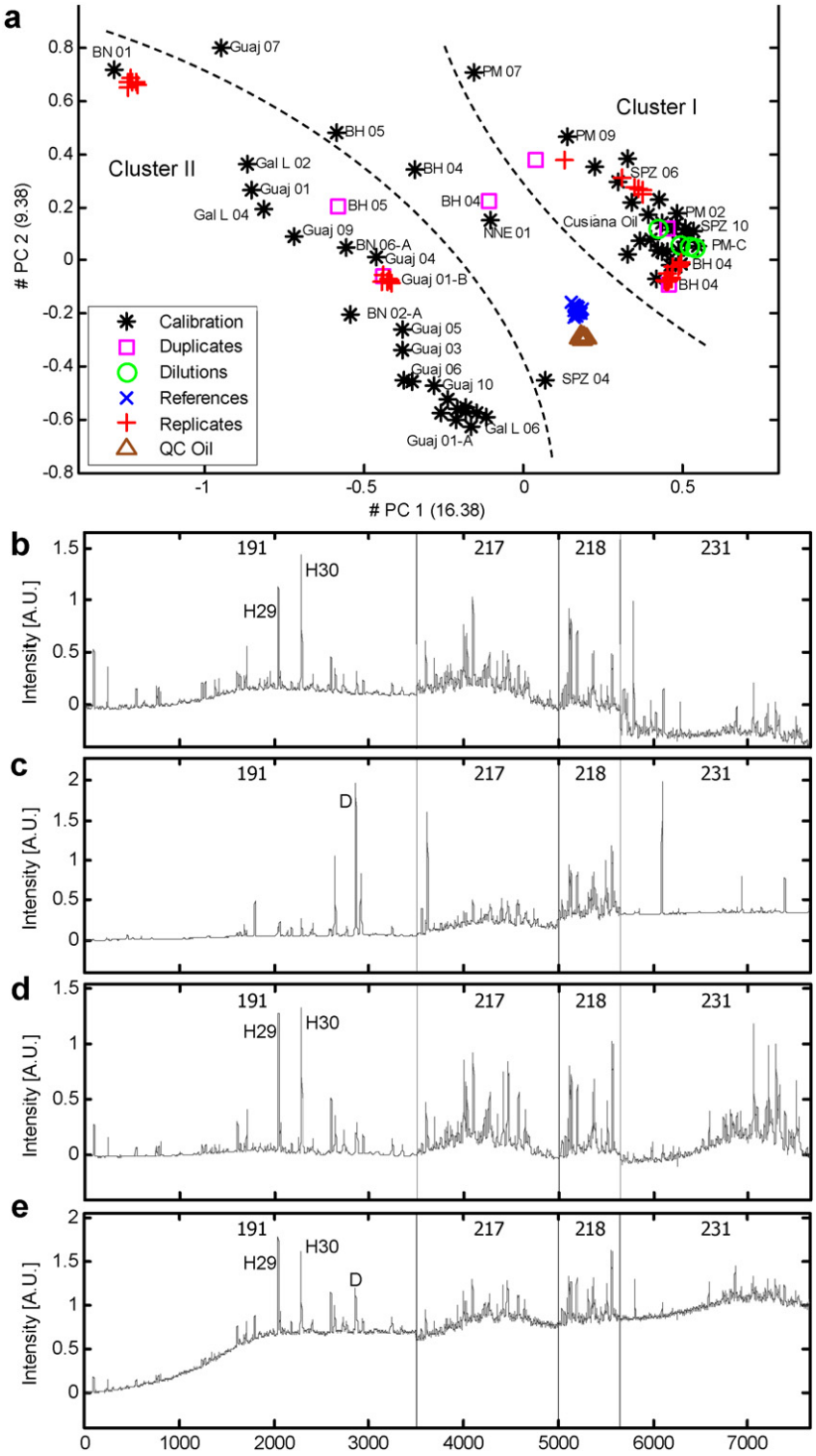

Fig. 6. Biomarkers data analysis: (a) Score plot of PC1 vs. PC2 for the principal component model of normalisation **Scheme III** (normalisation to Euclidean norm within SICs), and normalised data (**Scheme III**) of selected samples: Cluster I: (b) Cusiana oil; Cluster II: (c) BN 01, (d) Guaj 01-A and (e) Guaj 09. The symbols stand for: diploptene (D), 17α(H),21β(H)-30-norhopane (H29) and 17α(H),21β(H)-hopane (H30). Some labels were omitted from the Score plot (a) to improve the readability.

concentration of steranes and triaromatic steranes (large positive PC1 loading coefficients – Fig. S5) and a low relative concentration of vegetation biomarkers from recent organic matter, e.g. diploptene tentatively identified from literature [26] (Fig. S5 and Fig. 6c: BN 01). These latter peaks have negative PC1 loading coefficients (Fig. S5).

Most of the samples in Cluster II contain petroleum biomarkers as well (Fig. 6d and e), however with a different composition than samples from Cluster I. Specifically, the ratio of H29 and H30 vary in samples from Cluster II. Data of the quality control mixture of HFO from the Baltic Carrier and North Sea crude oil from the Brent oil field, included as an extra validation set ('QC Oil', marked

with '△' in Fig. 6a), plot in between the two clusters. The QC Oil is more related to the samples from outside the refinery than Cusiana oil.

4. Conclusion

The most contaminated samples are inside the refinery area. These samples present a petrogenic pattern and different weathering degrees. There are indications that both physical weathering and biodegradation have occurred. The former is related to the preferential loss of whole families of less alkylated PACs isomers, while the last is associated to changes in isomeric variations within alkylated PACs families in addition to loss of whole families. Samples from outside the refinery area are either less or not contaminated, or contain mixtures of diagenetic, pyrogenic and petrogenic inputs where different proportions predominate. The locations farthest away from industrial activity (e.g. Balsa Nova samples) contains, as expected, the lowest levels of PAC contamination. Regarding the biomarkers results, there are no evidences to conclude positive matches between the samples from outside the refinery area and the Cusiana oil. Using PCA of the pre-processed SICs is a step forward on source identification as unknown or not traditionally analysed compounds and unique chemical features are retained and provide more information for separating samples with similar hydrocarbon composition.

Acknowledgements

The study was financially supported by the Torkil Holms Foundation, the Lundbeck Foundation, the COWI Foundation and Petroleo Brasileiro S.A. We want to thank Rafael André Lourenço and Leandro Franco Macena de Araújo for their fundamental work on sample preparation, Jette Petersen for laboratory assistance, Leandro Rodrigues de Freitas e Ivanil Ribeiro Cruz for their assistance on mapping the study area, and Maria de Fátima G. Meniconi and Irene T. Gabardo for their constructive comments on the manuscript. We would also like to acknowledge Giorgio Tomasi for development of m-files used in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2012.02.041.

References

[1] J.W. Readman, G. Fillmann, I. Tolosa, J. Bartocci, J.-P. Villeneuve, C. Catinni, L.D. Mee, Mar. Pollut. Bull. 44 (2002) 48.
[2] UNEP/IOC/IAEA, Determination of petroleum hydrocarbons in sediments, Reference Methods for Marine Pollution Studies 20, 1992.
[3] S.G. Wakeham, C. Schaffner, W. Giger, Geochim. Cosmochim. Acta 44 (1980) 415.
[4] M.B. Yunker, R.W. Macdonald, R. Brewer, R. Vingarzan, R.H. Mitchell, D. Goyette, S. Sylvestre, Org. Geochem. 33 (2002) 489.
[5] American Society for Testing and Materials (ASTM), Standard Practice D 5739-06: Oil spill source identification by gas chromatography and positive ion electron impact low resolution mass spectrometry, 2006.
[6] J.H. Christensen, A.B. Hansen, G. Tomasi, J. Mortensen, O. Andersen, Environ. Sci. Technol. 38 (2004) 2912.
[7] Z. Wang, S.A. Stout, Oil Spill Environmental Forensics: Fingerprinting and Source Identification, Elsevier Academic Press, San Diego, 2007.
[8] Z. Wang, M. Fingas, D.S. Page, J. Chromatogr. A 843 (1999) 369.
[9] European Committee for Standardization (CEN), Oil Spill Identification – Waterborne petroleum and petroleum products – Part 2: Analytical methodology and interpretation of results, TC/BT TF 120 WI CSS27003, 2006.
[10] S.A. Stout, Z. Wang, in: R.E. Hester, R.M. Harrison (Eds.), Environmental Forensics, Royal Soc. Chem., Issues in Environmental Science and Technology, Special Publ. No. 26, London, 2008, p. 54.
[11] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Environ. Forensics 7 (2006) 77.
[12] M.F.G. Meniconi, S.M. Barbanti, in: Z. Wang, S.A. Stout (Eds.), Oil Spill Environmental Forensics: Fingerprinting and Source Identification, Elsevier Academic Press, San Diego, 2007, p. 505.
[13] S.A. Stout, A.D. Uhler, K.J. McCarthy, Environ. Forensics 2 (2001) 87.
[14] J.H. Christensen, A.B. Hansen, U. Karlson, J. Mortensen, O. Andersen, J. Chromatogr. A 1090 (2005) 133.
[15] J.H. Christensen, G. Tomasi, J. Chromatogr. A 1169 (2007) 1.
[16] J.H. Christensen, G. Tomasi, A.L. Scofield, M.F.G. Meniconi, Environ. Pollut. 158 (2010) 3290.
[17] Instituto Brasileiro de Geografia e Estatística (IBGE), Ministério do Planejamento, Orçamento e Gestão. Diretoria de Geociências. Coordenação de Recursos Naturais e Estudos Ambientais. Coordenação de Geografia. Indicadores de Desenvolvimento Sustentável Brasil 2008. (Série Estudos e Pesquisas. Informação Geográfica número 5). Dimensão ambiental – Água doce. Qualidade de águas interiores, cap. 11, Rio de Janeiro, 2008.
[18] M.F.G. Meniconi, I.T. Gabardo, M.E.R. Carneiro, S.M. Barbanti, G.C. Silva, C.G. Massone, Environ. Forensics 3 (2002) 303.
[19] HYDROGÉO PLUS INC., Avaliação dos Dados Ambientais de 2007–2008 e Proposições Futuras OSPAR/REPAR. Araucária, PR, Brasil, 2009.
[20] HYDROGÉO PLUS INC., Avaliação dos Dados Ambientais das Áreas Externas 2007–2009 OSPAR/REPAR, 2010.
[21] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[22] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[23] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemom. 20 (2006) 484.
[24] P. Baumard, H. Budzinski, Q. Michon, P. Garrigues, T. Burgeot, J. Bellocq, Estuar. Coast. Shelf Sci. 47 (1998) 77.
[25] Z. Wang, M. Fingas, S. Blenkinsopp, G. Sergy, M. Landriault, L. Sigouin, J. Foght, K. Semple, D.W.S. Westlake, J. Chromatogr. A 809 (1998) 89.
[26] M.I. Venkatesan, E. Ruth, P.S. Rao, B.N. Nath, B.R. Rao, Appl. Geochem. 18 (2003) 845.