Journal of Chromatography A, 1217 (2010) 8279–8289

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Selecting a reduced suite of diagnostic ratios calculated between petroleum biomarkers and polycyclic aromatic hydrocarbons to characterize a set of crude oils

R. Fernández-Varela, J.M. Andrade*, S. Muniategui, D. Prada

Dept. Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 1 July 2010
Received in revised form
27 September 2010
Accepted 11 October 2010

Keywords:
Crude oil
Biomarkers
Diagnostic ratios
SOMs
Heatmap
Procrustes Rotation

ABSTRACT

A set of 34 crude oils was analysed by GC–MS (SIM mode) and a suite of 28 diagnostic ratios (DR) calculated. They involved 18 ratios between biomarker molecules (hopanes, steranes, diasteranes and triaromatic steroids) and 10 quotients between polycyclic aromatic hydrocarbons. Three unsupervised pattern recognition techniques (i.e., principal components analysis, heatmap hierarchical cluster analysis and Kohonen neural networks) were employed to evaluate the final dataset and, thus, ascertain whether the crude oils grouped as a function of their geographical origin. In addition, an objective variable selection procedure based on Procrustes Rotation was undertaken to select a reduced set of DR that comprised for most of the information in the original data without loosing relevant information. A reduced set of four DR (namely; TA21, D2/P2, D3/P3 and B(a)F/4-Mpy) demonstrated to be sufficient to characterize the crude oils and the groups they formed.

1. Introduction

As industrialization progresses worldwide, more petroleum resources are required. Since demand increases and prices raise and they depend partly on the origin of the oil, there is a need for reliable analytical methodologies to screen whether a given oil batch corresponds to the contract agreement. It follows that chemical characterization of raw crude oils is a hot ongoing topic because of the strong efforts petrochemical companies undergone to exploit current productive extraction fields to their utmost. Disappointingly, chemical characterization is quite complex because the crude oil extracted at different wells in a same production field may be different. Even within a well the crude oil collected at different depths may be different because of the different mixture ratios than can occur within the source rocks [1].

The so-called petroleum biomarkers (or biological markers) are particularly useful to objectively assess the provenance of an oil [2] and their analysis was recommended to derive important information on the crude oils (oil fingerprinting), to differentiate oils, to search for the source of a spillage and to monitor the degradation process and the weathering stage of an oil under a variety of conditions [3]. Biomarkers are naturally occurring, ubiquitous and stable hydrocarbons that appear in crude oils and most petroleum products [2]. They derive from formerly living organisms whose organic materials were preserved in oil source rocks that upon burial (heat and pressure) generated crude oil over geologic time.

In this work hopanes, steranes, diasteranes and triaromatic steroids were analysed and they will be introduced briefly. In general, terpanes and steranes are branched cycloalkanes consisting of multiple condensed five- or six-carbon rings [4]. Hopanes are pentacyclic triterpanes which contain 27–35 carbon atoms. They derive from bacterial (prokaryotic) membrane lipid precursors. There, cyclization of squalene precursors give rise to hopanoids as, for instance, bacteriohopanetetrol [5]. Most hopanes derive from the most abundant one, C35 tetrahydroxybacteriohopane [6]. Steranes, formally named perhydrocyclopentanophenanthrene rings, are a class of 4-cyclic compounds derived from steroids or sterols (which constitute half of the lipids in the lipid membranes in all eukaryotic cells) via diagenetic and catagenetic degradation and saturation. The relative abundances of C27-, C28- and C29-steranes in oils reflect the carbon number distribution of the sterols in the organic matter in the source rocks [2]. Diasteranes are rearranged steranes that have no biological precursors, and are most likely formed during diagenesis and catagenesis. Triaromatic steroids can originate by aromatization and loss of a methyl group from monoaromatic steroids which, in turn, derived exclusively from sterols with a side-chain double bond during early diagenesis [5].

* Corresponding author. Tel.: +34 981 167000; fax: +34 981 167065.
E-mail address: andrade@udc.es (J.M. Andrade).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.10.043

Despite biomarkers can be used as such it was found more useful to calculate ratios among them [2,7,8], which were termed diagnostic ratios (DR). Besides, diagnostic ratios are unaffected by short-term weathering processes as long as they are based on compounds with little or comparable susceptibility to weathering. Some of them are known to relate to the thermal maturity of the source rocks that gave rise to a crude oil; the type of organic matter that was present in the source rock (e.g., terrestrial vs. marine); and/or the degree of weathering that may have occurred in the crude oil reservoir before oil extraction. Their specificity, diversity, complexity, and relative resistance to weathering make them useful 'markers' in the characterization of spilled oils, candidate source oils, and background contamination [9]. Understanding the meaning of these ratios, and evaluating whether they can be of use to characterize different oil production areas is therefore a must for petrochemical companies and researchers on environmental characterization of spilled hydrocarbons [2,9–11].

The diagnostic ratios considered in this study were based upon '*genetically*' significant (source-specific) variables that are known to occur among crude oils from different geologic basins as well as on several previous studies from different research groups [1,2,7–10,12–16].

In addition, the large amount of data generated in many environmental studies and/or petroleum quests require the use of multivariate chemometric tools to provide a unbiased, objective and defensible means to differentiate among qualitatively similar oils [9,10,17,18]. Chemometric approaches are also required to extract a reduced set of DRs which may suffice to still differentiate among the different types of oils.

In this paper 40 biomarkers and 16 polycyclic aromatic hydrocarbons (PAHs) were analysed in 34 crude oils by gas chromatography with mass spectrometry detection (GC–MS), using the SIM mode. In total, 28 DRs were calculated and studied by unsupervised pattern recognition techniques (principal components analysis, heatmap hierchical cluster analysis and Kohonen neural networks) in order to assess whether they may differentiate different petroleum-producing geographical areas worldwide. Further, an objective variable selection technique based on Procrustes Rotation was applied to extract a minimum set of relevant DRs to differentiate among the different petroleum basins as much as possible. Procrustes Rotation was selected because of its ability to select analytical variables instead of abstract factors or combinations of variables like other chemometric techniques do.

2. Experimental

2.1. Samples

Thirty-four crude oil samples representing different petroleum bearing basins throughout the world were collected from several Spanish refineries (Table 1). Their specific gravities ranged from 19° to 48° API. Crude oil samples were water and sediment extracted following an ASTM guide [19], light-protected and stored at 4 °C until analysis. The analytes were measured after dissolving 20–50 mg of each sample, weighted accurately in an analytical balance, in 5 mL of dichloromethane (Super purity solvent, Merck).

2.2. Gas chromatography–mass spectrometry

An HP 6890 instrument (Agilent Technologies, Palo Alto, CA, USA) with a pulsed splitless injector, an HP 5973 mass spectrometry detector and an HP-5MS fused silica capillary column (J&W Scientific, Folsom, CA, USA) 60 m long (0.25 mm i.d., 0.25 μm film thickness) were employed. Operating conditions were: starting oven temperature, 40 °C, held isothermally for 1 min, and raised to 300 °C at 6 °C/min and held isothermally for 30 min. Carrier gas: Helium, 1 mL/min constant flow. Injector and transfer line temperature were 300 and 280 °C, respectively. Ionization energy: 70 eV, ion source temperature, 230 °C. Injection was performed in the pulsed splitless mode, injected sample: 1 μL. The *m*/*z* range for MS analysis was 40–440. The SIM mode (selected ion monitoring) was used throughout. In total, 20 hopanes, 13 steranes and diasteranes, 7 triaromatic steroids biomarkers and 16 PAHs were analysed (see Table 2 for more details).

Table 1
Resume of the crude oils employed in this study.

Origin		Product
North Africa	Libya	Amna, Es Sharara, Sarir, Sirtica
	Algeria	Sahara Blend
	Tunisia	Ashtart
Middle East	Azerbaijan	Azeri Light
	Iran	Foroozan, Soroosh
	Saudi Arabia	Arabian Heavy
	Syria	Syria
Central Africa	Nigeria	Brass, Ea, Escravos
	Ecuatorial Guinea	Zafiro
North Sea	North Sea	Brent, Draugen, Ekofisk, Flotta, Forties, Gullfaks, Norne, Schiehallion, Statfjord
South America	Argentina	Cañadón Seco
	Colombia	Caño Limón, Vasconia
	Ecuador	Oriente
	Venezuela	Santa Barbara
South Africa	Angola	Girassol
Central America	México	Maya
Russia	Russia	Siberian Light, Tengiz, Ural Light

As mentioned in the introduction, different diagnostic ratios (DRs) have been proposed in literature to differentiate crude oils and the most common ones were selected to perform this work: 27Ts, 28ab, 25nor30ab, 29Ts, 30O, 30G, 29ab, 30d, 32abS, 27dia, 29aaS, 29bb, 27bbSTER, 28bbSTER, 29bbSTER, TA21, TA26, TA27, D2/P2, D3/P3, D3/C3 and Retene/P4 [20]. Besides, common diagnostic PAHs include dibenzothiophenes and phenanthrenes, although some other possibilities exist [13,21,22]. Thus, 'source-specific' marker compounds, including alkylated PAH hydrocarbons within homologous alkylation isomeric groups were identified as well and their ratios calculated. 'Source-specific' here means that the DRs may serve as unambiguous markers for some oils under study, as many times they are subject to little interference from absolute concentration fluctuation of individual compounds [15]. The 'source-specific' DRs considered here were 2-MP/1-MP and 4-MD/1-MD [20], B(a)F/4-Mpy, B(b+c)F/4-Mpy, 2-Mpy/4-Mpy and 1-Mpy/4-Mpy [23].

Further, in order to calculate the DRs a previous internal quality control evaluation was done, as it had been shown that biomarkers may be affected by the analytical variability and sample heterogeneity [8,20]. All samples were analysed by triplicate and the relative standard deviation (RSD) of each compound calculated. Then, following [20] and [23], only DRs for which the RSDs of the compounds involved were lower than 5% were employed. Accordingly, a suite of 28 DR (18 quotients between the peak heights of several biomarkers and 10 ratios between peak areas for several PAHs) were calculated. Their full description is displayed in Table 3.

2.3. Chemometric techniques and software

Here unsupervised pattern recognition multivariate techniques had to be used because the lack of more samples of known origin impeded us to get independent validation sets of samples that might be used to fully validate supervised methods. Hence, three unsupervised methods were selected. Two of them, principal com-

Table 2
Description of the biomarker compounds analysed in this work.

Group	Name	Abbreviation	*m/z*
Biomarkers			
Hopanes	18α(H)-2,29,30-trisnorhopane	27Ts	191
	17α(H)-22,29,30-trisnorhopane	27Tm	191
	17α(H),21β(H)-28,30-bisnorhopane	28ab	191
	17α(H),21β(H)-25-norhopane	25nor30ab	191
	17α(H),21β(H)-30-norhopane	29ab	191
	18α(H)-30-norneohopane	29Ts	191
	15α-metil-17α(H)-27-norhopane (diahopane)	30d	191
	17α(H),21α(H)-30-norhopane(normoretane)	29ba	191
	18α(H)-oleanane	30O	191
	17α(H),21β(H)-hopano	30ab	191
	17α(H),21α(H)-hopane(moretane)	30ba	191
	17α(H),21β(H), 22S-homohopane	31abS	191
	17α(H),21β(H), 22R-homohopane	31abR	191
	Gammacerane	30G	191
	17α(H),21β(H), 22S-bishomohopane	32abS	191
	17α(H),21β(H), 22R-bishomohopane	32abR	191
	17α(H),21β(H), 22S-trishomohopane	33abS	191
	17α(H),21β(H), 22R-trishomohopane	33abR	191
	17α(H),21β(H), 22S-tetrakishomohopane	34abS	191
	17α(H),21β(H), 22R-tetrakishomohopane	34abR	191
Steranes and diasteranes	13β(H),17α(H), 20S-cholestane (diasterane)	27dbS	217
	13β(H),17α(H), 20R-cholestane (diasterane)	27dbR	217
	24-metil-5α(H),14α(H),17α, 20R-cholestane	28aaR	217
	24-etil-5α(H),14α(H),17α, 20S-cholestane	29aaS	217
	24-etil-5α(H),14β(H),17β, 20S-cholestane	29bbS	217
	24-etil-5α(H),14α(H),17α, 20R-cholestane	29aaR	217
	5α(H),14β(H),17β(H), 20R-cholestane	27bbR	218
	5α(H),14β(H),17β(H), 20S-cholestane	27bbS	218
	24-metil-5α(H),14β(H),17β(H), 20R-cholestane	28bbR	218
	24-metil-5α(H),14β(H),17β(H), 20S-cholestane	28bbS	218
	24-etil-5α(H),14β(H),17β(H), 20R-cholestane	29bbR	218
	24-etil-5α(H),14β(H),17β(H), 20S-cholestane	29bbS	218
	24-etil-5α(H),14β(H),17β(H), 20R-cholestane	29bbR	218
Triaromatic steroid	C20-triaromatic steroid hydrocarbon	C20TA	231
	C21-triaromatic steroid hydrocarbon	C21TA	231
	C26, 20S-triaromatic steroid hydrocarbon	SC26TA	231
	C26, 20R + C27, 20S-triaromatic steroid hydrocarbon	RC26TA + SC27TA	231
	C28, 20S-triaromatic steroid hydrocarbon	SC28TA	231
	C27, 20R-triaromatic steroid hydrocarbon	RC27TA	231
	C28, 20R-triaromatic steroid hydrocarbon	RC28TA	231
PAHs	C_2-phenanthrene	P2	206
	C_3-phenanthrene	P3	220
	C_4-phenanthrene	P4	234
	1-Methylphenanthrene	1MP	192
	2-Methylphenanthrene	2MP	192
	Retene	Retene	234
	C_2-dibenthiophene	D2	212
	C_3-dibenzothiophene	D3	226
	1-Methyldibenzothiophene	1MD	198
	4-Methyldibenzothiophene	4MD	198
	C_3-chrysene	C3	270
	Benzo(a)fluorene	B(a)F	216
	Benzo(b + c)fluorene	B(b + c)F	216
	1-Methylpyrene	1-Mpy	216
	2-Methylpyrene	2-Mpy	216
	4-Methylpyrene	4-Mpy	216

ponents analsis – PCA, and hierarchical cluster analysis – CA, are parametric classical methods with proved abilities to unravel the main patterns within the datasets. The third one, Kohonen neural networks or self-organizing maps – SOM, is a so-called natural computation technique as it uses rules based on how humans process information rather than formal equations to get the model. Slight differences in the groups yielded by these techniques are expected because their fundamentals are quite different. The relevant point here is that as long as the major results are not different, the final conclusions are supported by different methodologies and, thus, became trustworthy and somehow 'validated'. In this particular study, as PCA takes into account the relationships between the variables and discard noisy information (which becomes relegated to the last PCs), it will be the 'reference method'. On the contrary, hierarchical clustering may be affected either by the correlation between the variables and/or the presence of noisy or unrelevant information (which is not known *a priori*). This, in turns, justifies the usage of a variable selection procedure.

PCA aims to reduce dimensionality of the problem in the DRs domain. This allows studying correlations between the variables and defining some new factors that comprise the most relevant information. By definition, the first factors (PCs) explain more information than the latter ones. Also of relevant importance is the capability of PCA to show how the samples look like in the experimental domain and, thus, insightful and very accurate conclusions can be drawn regarding the samples (and the variables).

Hierarchical clustering looks for groups of samples (clusters) so that the samples within them are similar and, at the same time, the

Table 3
Details of the diagnostic ratios calculated using PAHs and biomarkers. Acronyms used for the biomarkers correspond to those in Ref. [23] whereas those for PAHs correspond to Refs. [13,23].

Biomarkers – hopanes	
27Ts	[27Ts(191)]/([27Ts(191)] + [27Tm(191)]) * 100
28ab	[28ab(191)]/([28ab(191)] + [30ab(191)]) * 100
25nor30ab	[25nor30ab(191)]/([25nor30ab(191)] + [30ab(191)]) * 100
29Ts	[29Ts(191)]/([29Ts(191)] + [30ab(191)]) * 100
30O°	[30O(191)]/([30O(191)] + [30ab(191)]) * 100
30G	[30G(191)]/([30G(191)] + [30ab(191)]) * 100
29ab	[29ab(191)]/([29ab(191)] + [30ab(191)]) * 100
30d	[30d(191)]/([30d(191)] + [30ab(191)]) * 100
32abS	[32abS(191)]/([32abS(191)] + [32abR(191)]) * 100
Biomarkers – steranes and diasteranes	
27dia	([27dbS(217)] + [27dbR(217)])/([27dbS(217)] + [27dbR(217)] + [27bbR(217)] + [27bbS(217)]) * 100
29aaS	[29aaS(217)]/([29aaS(217)] + [29aaR(217)]) * 100
29bb	([29bbR(217)] + [29bbR(217)])/([29bbS(217)] + [29bbR(217)] + [29aaR(217)] + [29aaS(217)]) * 100
27bbSTER	[27bb(S + R)(218)]/([27bb(S + R)(218)] + [28bb(S + R)(218)] + [29bb(S + R)(218)]) * 100
28bbSTER	[28bb(S + R)(218)]/([27bb(S + R)(218)] + [28bb(S + R)(218)] + [29bb(S + R)(218)]) * 100
29bbSTER	[29bb(S + R)(218)]/([27bb(S + R)(218)] + [28bb(S + R)(218)] + [29bb(S + R)(218)]) * 100
Biomarkers – triaromatic steroids	
TA21	[C21TA(231)]/([C21TA(231)] + [RC28TA(231)]) * 100
TA26	[SC26TA(231)]/([SC26TA(231)] + [SC28TA(231)]) * 100
TA27	[RC27TA(231)]/([RC27TA(231)] + [RC28TA(231)]) * 100
PAHs	
D2/P2	[C_2-dibenzothiophenes/(C_2-dibenzothiophenes + C_2-phenanthrenes)] * 100
D3/P3	[C_3-dibenzothiophenes/(C_3-dibenzothiophenes + C_3-phenanthrenes)] * 100
D3/C3	[C_3-dibenzothiophenes/(C_3-dibenzothiophenes + C_3-chrysenes)] * 100
2-MP/1-MP	[2-methylphenanthrene/(2-methylphenanthrene + 1-methylphenanthrene)] * 100
4-MD/1-MD	[4-methyldibenzothiopene/(4-methyldibenzothiophene + 1-methyldibenzothiophene)] * 100
Retene/P4	[retene/(retene + C_4-phenanthrenes)] * 100
B(a)F/4-Mpy	[benzo(a)fluorene/(benzo(a)fluorene + 4-methylpyrene)] * 100
B(b + c)F/4-Mpy	[benzo(b + c)fluorene/(benzo(b + c)fluorene + 4-methylpyrene)] * 100
2-Mpy/4-Mpy	[2-methylpyrene/(2-methylpyrene + 4-methylpyrene)] * 100
1-Mpy/4-Mpy	[1-methylpyrene/(1-methylpyrene + 4-methylpyrene)] * 100

groups can be differentiated among them. Two parameters need to be defined: a similarity criterion, by using a metric of the distance between the samples (there are many possibilities), and a clustering rule to proceed with the formation of the groups (there are different possibilities also).

In this paper the so-called 'heatmap' hierarchical clustering was applied. This way of presenting the results is unusual in the environmental field (despite it has proved valuable in other scientific fields, like in genetic studies) and was applied because it is highly useful to interpret chemically why some groups of samples appeared. The heatmap is a colour-coded two-dimensional mosaic formed by the joint representation of two clusters, one is sample-oriented while the other is variable-oriented. Thus, the heatmap describes all similarities within the whole dataset (samples vs. diagnostic ratios). Note that, as in any typical clustering, the two parameters (similarity and clustering method) need to be selected for each of the two dendrograms. Despite any selection is mathematically correct some trials must be performed to select those parameters that yield the most chemically interpretable groups. This is so because the different distances and/or clustering methods may lead to different groups of samples. As in the heatmap two dendrograms are developed, the trials are a bit more time-consuming than in common clustering.

Each tile of the mosaic is coloured with a different intensity according to the values of the (pre-processed) data. Hence, the heatmap literally adds another dimension of information presented by the dendrograms, which may facilitate its interpretation [24].

Artificial neural networks based on the Kohonen approach (Kohonen Maps or SOMs) are self-organising systems capable of solving unsupervised problems. In SOMs similar input objects are linked to the topologically closest neurons in the network, i.e., neurons that are located close to each other have similar reactions to similar inputs, while the neurons that are far apart have different reactions to similar inputs. The SOM map is characterized usually by being a squared toroidal space that consists of a grid of neurons (the 'topology'). Each neuron contains as many elements (weights) as the number of input variables. The weights of each neuron are randomly initialised between 0 and 1 and updated on the basis of the input vectors (i.e., samples for a certain number of times (called training epochs). Both the number of neurons and epochs to be used to train the map must be defined by the user; more technical details can be found elsewhere (e.g., [25,26]).

Objective variable selection can be performed in different ways although the Procrustes Rotation (PR) technique was employed here. Its conceptual ideas and general applications were discussed elsewhere [27,28] and they will not be included here. The first step for variable selection is to determine the optimum number of factors (principal components, PCs) that possess relevant information, rather than noise. A straightforward way to do so is to evaluate the amount of information lost when the variables are successively extracted for a given number of principal components. Irrelevant factors will be characterized by a stabilization on the loss of information regardless of the variable been deleted and a low value in the overall variance they explain [29,30]. Then, the variable selection proceeds by deleting each variable in turn and evaluating the information lost in each case [27,28].

The chemometric multivariate studies were performed using built-in and in-house routines for Matlab® (the Mathworks Inc., MA, USA) and GenEx® (MultiD Analysis, Göteborg, Sweden).

3. Results and discussion

3.1. GC–MS results and univariate studies

Fig. 1 depicts the general appearance of the fragmentograms of the *m*/*z* 191, 217, 218 and 231 ion profiles where from the DRs were calculated for the crude oils.

A traditional univariate study for the experimental values of the 28 DRs calculated for all products and a classical representation of their distribution [20] showed only that some ratios presented

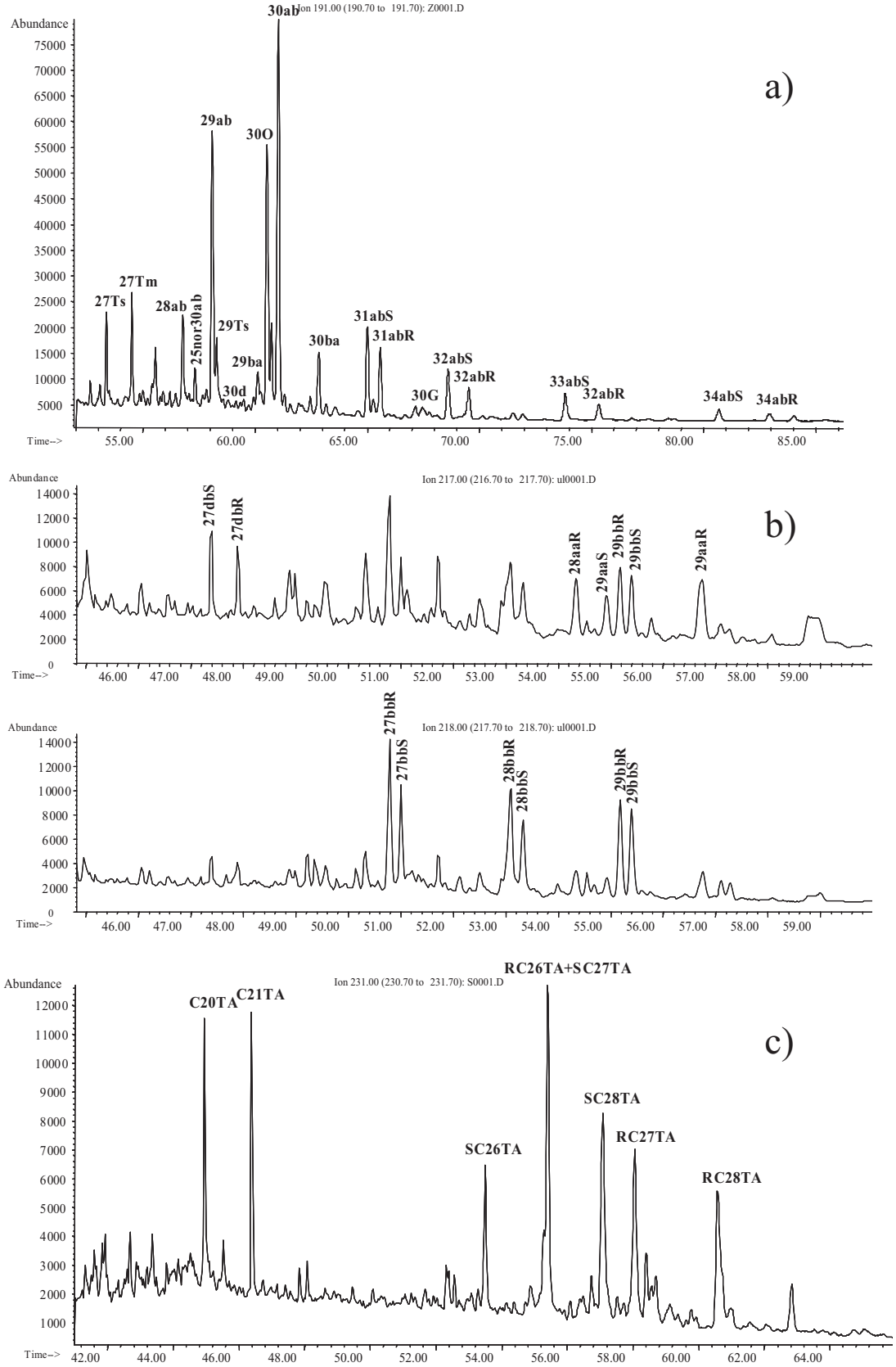

Fig. 1. Representative GC–MS fragmentograms of the studied biomarkers for the crude oils: (a) hopanes (*m*/*z* 191), (b) steranes (*m*/*z* 217) and diasteranes (*m*/*z* 218), and (c) triaromatic steroids (*m*/*z* 231).

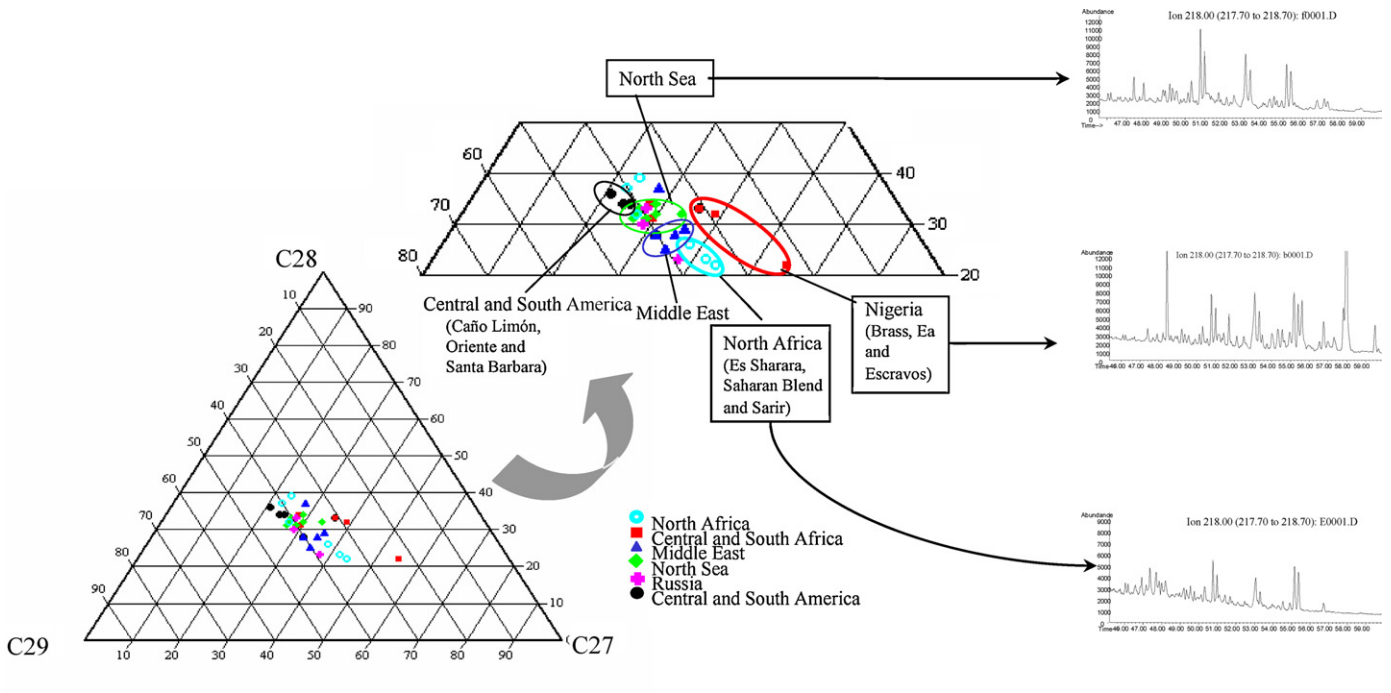

Fig. 2. Ternary plot of C_{27}, C_{28} and C_{29} steroid abundance for the crude oils dataset.

large variability and some clustering, which suggested that they may be of help to differentiate some crude oils (e.g., 27Ts, 27dia, D2/P2 and D3/P3). On the contrary, most DRs seemed not suitable for differentiation purposes, like 30G, 30d and 32abS.

To extract more information from the raw data, ternary diagrams were proposed some time ago to represent the C_{27}, C_{28} and C_{29} steranes [2,31–33]. Here the ternary diagram revealed some differences among the samples (Fig. 2) and it differentiated quite well some (but not all) North African and South American oils. Unfortunately, the groups were not perfect.

The Nigerian crude oils (Brass, Ea and Escravos) had similar sterane distributions, which suggested a same origin for the parent organic matter. Noteworthy, other North African crude oils (Amna, Ashtart and Sirtica) situated far from them. Southern and Central American oils (Colombia, Ecuador and Venezuela) became clustered at the left upper part of the graph, characterized by highest values of C_{27} and lowest of C_{29}; Argentina and México oils were not included here.

Crude oils from the North Sea formed a quite homogenous group (although it included some other oils), with medium values for the three variables. The Middle East samples grouped very well but for Azeri Light, which situated at the top of the graph and they were characterized by highest values of C_{27} and lowest of C_{28}.

A final conclusion of the univariate studies was that no variable (not even classical subsets, as the ternary diagrams) could differentiate among the major oil production areas worldwide and, therefore, a multivariate approach seemed in order.

3.2. Principal components analysis

A principal components analysis (PCA) was made on the mean centred data. The first three PCs explained 89.1% of the total variance and they sufficed to describe the system. Fig. 3 revealed five groups of crude oils considering the sample scores. One appeared geographically linked to North Africa, and it was formed by the Es Sharara, Sahara Blend and Sarir crude oils. The Amna (Lybia) and Ashtart (Tunissia) oils were close to this group on the PC3 direction; despite they did not so in the PC1 and PC2 directions (graph not shown) and, so, it was decided not to include them in Group 1.

Group 2 was composed of almost all Middle East oils (Arabian Heavy, Foroozan, Syria and Soroosh) despite other heavy oils got also here; namely, Tengiz (Russia), Cañadón Seco (Argentina) and Maya (México). Group 3 was formed by crude oils from different origins: Central and South America, Middle East, and North Africa. They were Caño Limón (Colombia), Vasconia (Colombia), Oriente (Ecuador), Santa Barbara (Venezuela), Sirtica (Lybia), Girassol (Angola), Ural Light (Russia), Siberia Light (Russia) and Azeri Light (Azerbaijan). Group 4 was constituted by Central Africa oils: Brass, Ea, Escravos (all from Nigeria) and Zafiro (Equatorial Guinea). Group 5 grouped the Brent, Draugen, Ekofisk, Flotta, Forties, Gullfaks, Norne, Schiehallion and Statfjord oils, all from the North Sea.

The profile of the loadings for PC1 (50.8% of the overall variance) is dominated by D2/P2, D3/P3 and D3/C3 (positive loadings) and 27Ts, B(a)F/4-Mpy and B(b + c)F/4-Mpy (negative loadings). PC1 differentiated mostly Group 2 (Middle East oils) from other crude oils as they had maximum values for the ratios with positive loadings. For instance, 75, 79 and 97, for D2/P2, D3/P3 and D3/C3, for Arabian Heavy, vs. 16, 18 and 66 (respectively) for Central Africa crude oils (e.g., brass, in the opposite side of PC1).

PC2 (23.4% explained variance) was mainly defined by 28ab, 30O, TA26 and B(a)F/4-Mpy (positive loadings); and 27Ts and TA21 (negative loadings). PC2 opposes Group 1 (constituted by crude oils from North Africa: Es Sharara, Sahara Blend and Sarir; with minimum values on the diagnostic ratios associated to positive loadings and close-to-zero values for 28ab and 30O) to other oils, specially to group 4 (Central Africa crude oils, with medium-high values for 28ab, TA26 and B(a)F/4-Mpy and maximum values for 30O). Interestingly, the ratio 30O involving the oleanane molecule is characteristic of Central Africa Nigerian crude oils [34].

The most relevant variables for PC3 (14.9% variance) were 27Ts, 25nor30ab, 27dia, TA21, B(a)F/4-Mpy and B(b + c)F/4-Mpy (positive loadings) and 30O, 29ab, 29bbSTER, 2-MP/1-MP, 2-Mpy/4-Mpy and 1-Mpy/4-Mpy (negative loadings). PC3 was relevant to differentiate samples from the North Sea, termed Group 5, as they showed highest experimental ratios for the variables with highest positive

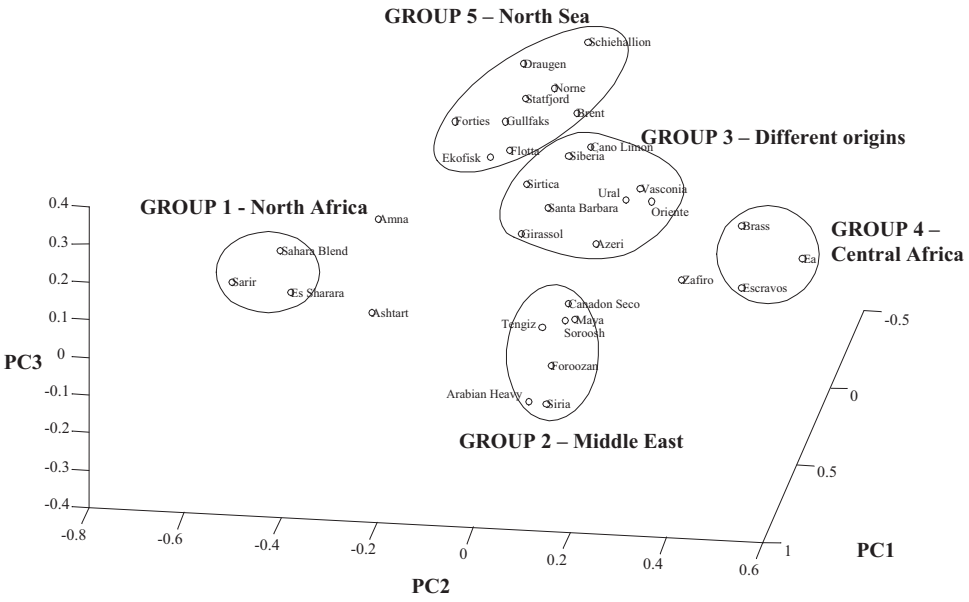

Fig. 3. PCA scores scatterplots (mean centred data) considering all variables for the crude oils dataset.

loadings (e.g., 61, 9, 56, 67, 70 and 54 for 27Ts, 25nor30ab, 27dia, TA21, B(a)F/4-Mpy and B(b + c)F/4-Mpy, respectively, for Schiehallion) opposed to Middle East crude oils, which presented lowest values for these variables (e.g., 33, 4, 16, 56, 22 and 7, respectively, for Syria).

3.3. Cluster analysis

The Manhattan distance and the Ward's clustering algorithm were used to cluster the variables whereas the Euclidean squared distance and Ward's clustering algorithm were considered to group the samples. Columnwise mean centred data (i.e., mean centring across the variables) was selected after some preliminary trials.

The heatmap yielded a combined dendrogram which differentiated clearly between four groups of samples (Fig. 4). They did not agree exactly with those obtained by PCA. Cluster 'A' matched with group 2 from PCA, but for Cañadón Seco (Argentina). This group, formed by Middle East oils, is characterized by highest values on several DR calculated between PAHs, as D2/P2, D3/P3 and D3/C3. Cluster 'B' agrees with group 1 of PCA, grouping North Africa crude oils, which got defined by highest values on TA21 and at the same time lowest values on TA27.

Cluster 'C' showed two distinctive behaviours and, accordingly, can be subdivided: subgroup C1 constituted by North Sea oils (corresponding to group 5 of PCA) and subgroup C2 linked to the Oriente, Santa Barbara, Siberian Light and Ural Light crude oils (which formed Group 3 in PCA—different origins). Subgroup C1 was defined by high (but no maximum) values for 28ab, B(b + c)F/4-Mpy and B(a)F/4-Mpy. Subgroup C2 was characterized by intermediate values of D2/P2, D3/P3 and D3/C3.

Cluster 'D' included two groups. Subgroup D3 was constituted by Nigerian crude oils (group 4 from PCA). They had highest values on 30O (a ratio involving 30-oleanane, which is present only in Nigerian crude oils). Subgroup D4 was linked to crude oils forming group 3 of PCA (Girassol, Sirtica, Caño Limón, Vasconia, Azeri Light) plus the Amna (located close to group 1 from PCA), Cañadón Seco (group 2 from PCA), Ekofisk (group 5) and Zafiro (group 4) oils. Subgroup D4 presented highest values on the TA26 ratio.

To sum up, groups A, B, C1 and D3 from hierarchical clustering had the same geographical interpretation as the corresponding groups from the PCA above. Main difference with the PCA studies was that samples that formed cluster D4 were located in different PCA groups. Nevertheless, in all cases those samples were not at the core of the PCA groups, but slightly apart (which pointed out that they were not too similar to the core of the groups after all); and this might be the reason why hierarchical clustering grouped them in a separate cluster.

3.4. Kohonen neural networks

In this work, several topologies for the SOMs were studied, from 7×7 to 12×12 (in all cases columnwise mean centred data were used). In addition, the number of neighbours varied from 6 to 7. The learning ratio and the number of iterations were varied as well. The final choice was a 7×7 map, 7 neighbours, a 0.3 learning ratio and 75 iterations.

Although the topology of the SOMs can be arranged in a cyclic lattice, we found out best results using non-cyclic topologies and these are presented here. Fig. 5 revealed that five distinct groups of crude oils appeared, very similar to those from PCA. Group 1 coincided with that from PCA. Group 2 at SOM was equal to group 2 from PCA but for Cañadón Seco (which locate now at group 3). Group 3 at SOM agreed with group 3 from PCA, although the Ashtart, Cañadón Seco and Flotta became included here, and the Sirtica oil was not considered. Group 4 from the SOM coincided with group 4 from PCA and it is formed by the Nigerian crude oils and Zafiro (Equatorial Guinea). This group is, hence, consituted by Central Africa oils. Group 5 was formed by North Sea oils, although the Amna and Sirtica oils (Lybia) got clustered here as well.

3.5. Variable selection

Fig. 6a presents the complex situation obtained in this study to select the optimum number of factors (PCs) to be used. Although it was clear that after 9 PCs, the factors appeared much less relevant than the first ones, a first minimum on the loss of information was also noticed at 4 PCs as well. Hence, 4 and 9 PCs were considered as candidates to represent the dataset. In case the results derived from them were similar, 4 PCs should be the choice due to parsimony. According to the PR algorithms the minimum number of variables to be retained coincides with the number of optimum components that describe optimally the system. Hence a minimum

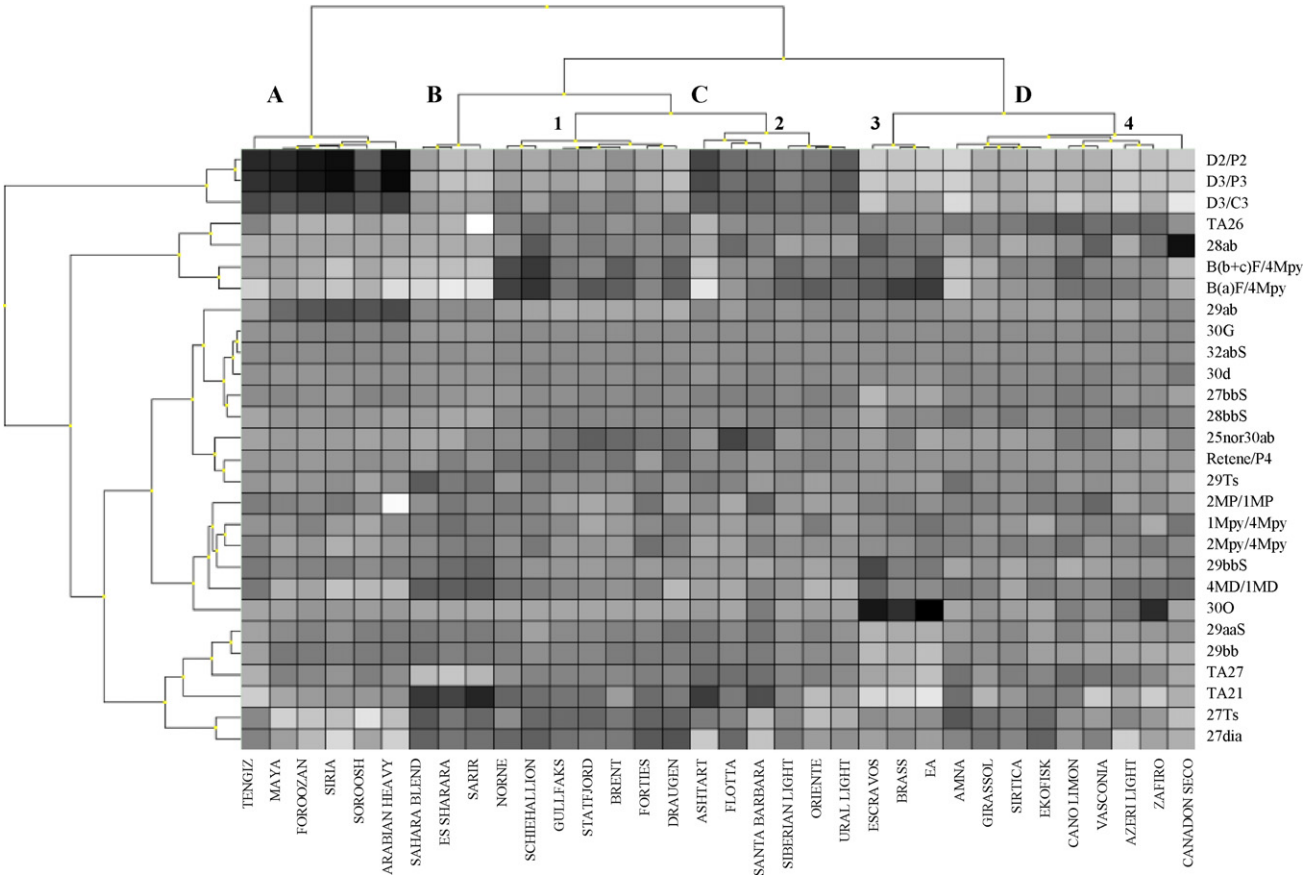

Fig. 4. Heatmap hierarchical clustering for the crude oils dataset using all the DRs. The experimental values of the DRs were coded with tiles whose shadow range from black (■, highest values) to white (□, lowest values).

set of four original DR were retained; namely TA21, D2/P2, D3/P3 and B(a)F/4-Mpy. Fig. 6b and c demonstrates that they satisfactorily reconstructed the distribution of the samples on the PC1–PC2 subspace. Results considering 9 PCs did not improved those with 4 PCs and, so, the latter were considered the final choice. It is worth noting that the four selected variables were highly relevant on the loadings defining the first two PCs at the PCA above considering all variables. On the contrary, 30O that was specific for Nigerian oils, was not selected. Those oils appear characterized now by highest values on B(a)F/4-Mpy and (at the same time) lowest values on TA21 (see discussions below). Also interesting, the DRs selected seem reasonable as the relative amounts of phenanthrenes and dibenzothiophenes had previously been considered useful for source identification and to assess the extent of oil weathering [13]. Particularly, D2/P2 and D3/P3 remained reasonably constant as the Exxon Valdez cargo degraded under the conditions of Prince William Sound [13].

Fig. 6c obtained from a PCA of the selected four variables revealed five groups of crude oils, which agreed with those derived

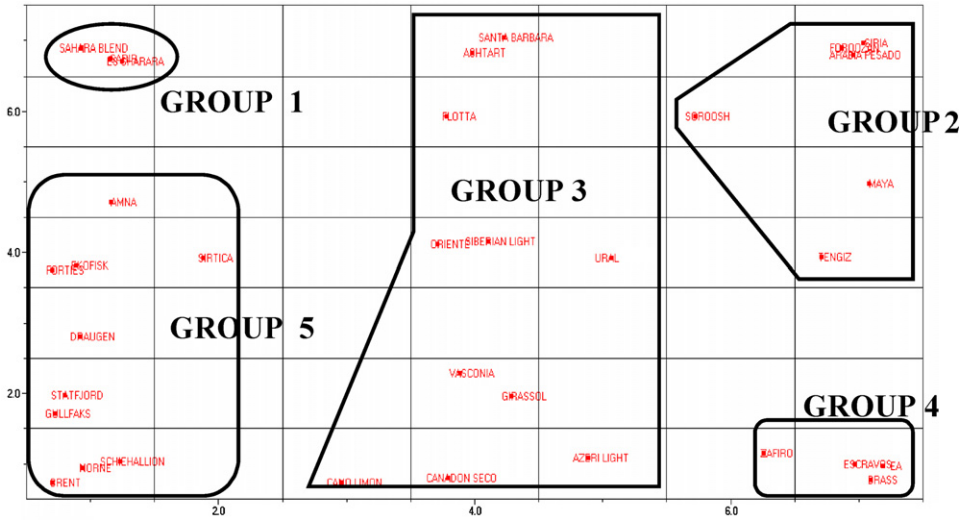

Fig. 5. SOM map for the crude oils dataset employing all the measured variables.

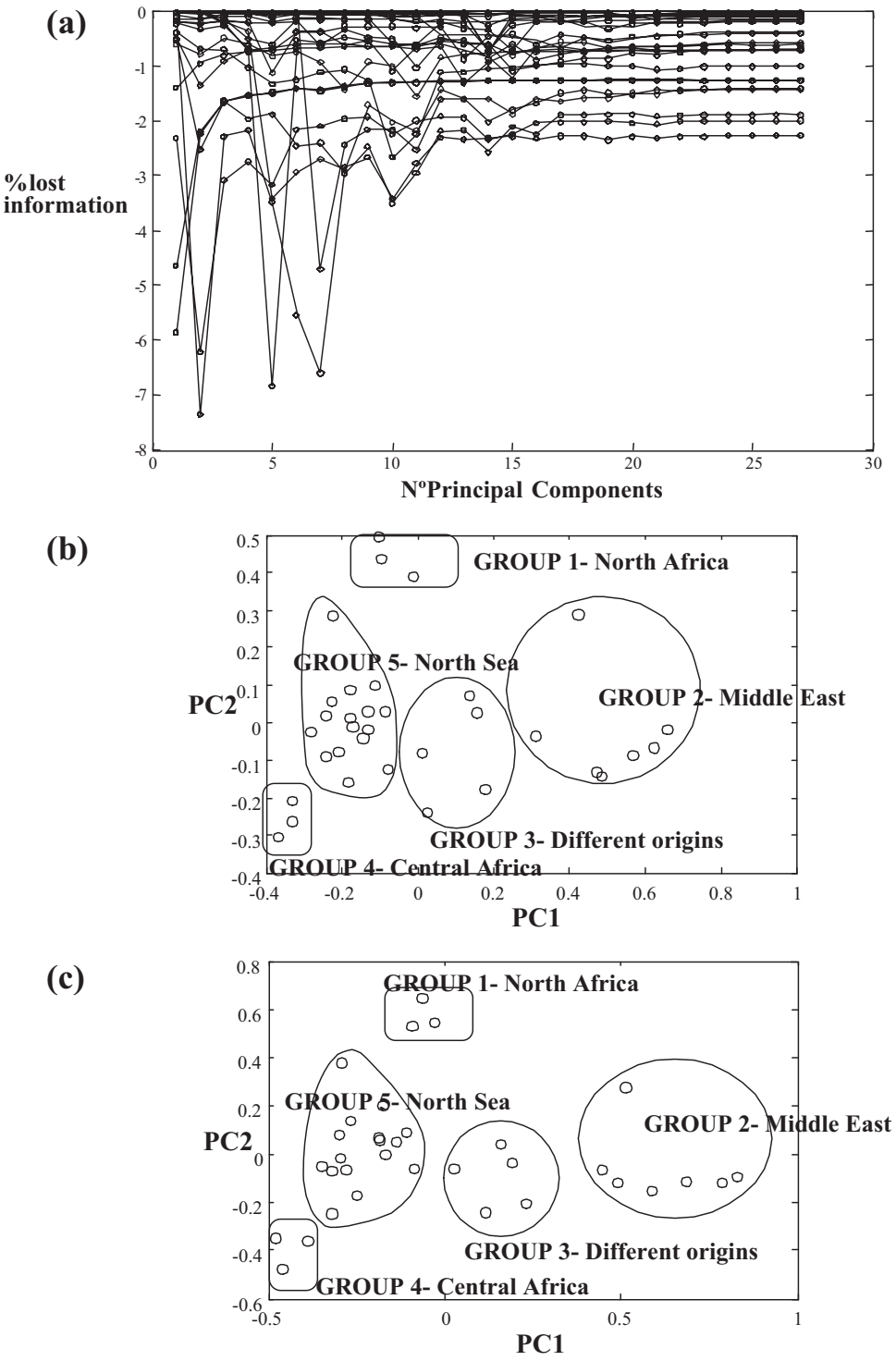

Fig. 6. (a) Amount of information lost when each variable is deleted in turn for each number of PCs to select the number of principal components and PCA scores scatterplots (mean centred data) considering: (b) all variables and (c) four variables.

from the PCA above employing all the DRs. Group 1 (constituted by crude oils from North Africa) presented maximum values for TA21, high values (but not maximum) for D2/P2 and D3/P3, and low values (but not minimum) for B(a)F/4-Mpy. Group 2 (Middle East crude oils) presented intermediate values for TA21, high values (maxima in some cases) on D2/P2 and D3/P3, and medium-low values for B(a)F/4-Mpy. Group 4 (Central Africa crude oils) showed intermediate-low values for TA21, lowest values for D2/P2 and D3/P3, and highest ones for B(a)F/4-Mpy. Finally, group 5, formed by North Sea crude oils, had highest values on TA21, intermediate-low values on D2/P2 and D3/P3, and intermediate-low values for B(a)F/4-Mpy.

The heatmap presented in Fig. 7 revealed clusters with slight differences regarding those considering all variables (Fig. 4). It was obtained using the Manhattan distance and the Ward's clustering algorithm to cluster the variables, and the Euclidean squared distance and the Ward's clustering algorithm to group the samples. Columnwise mean centred data was selected after some prelim-

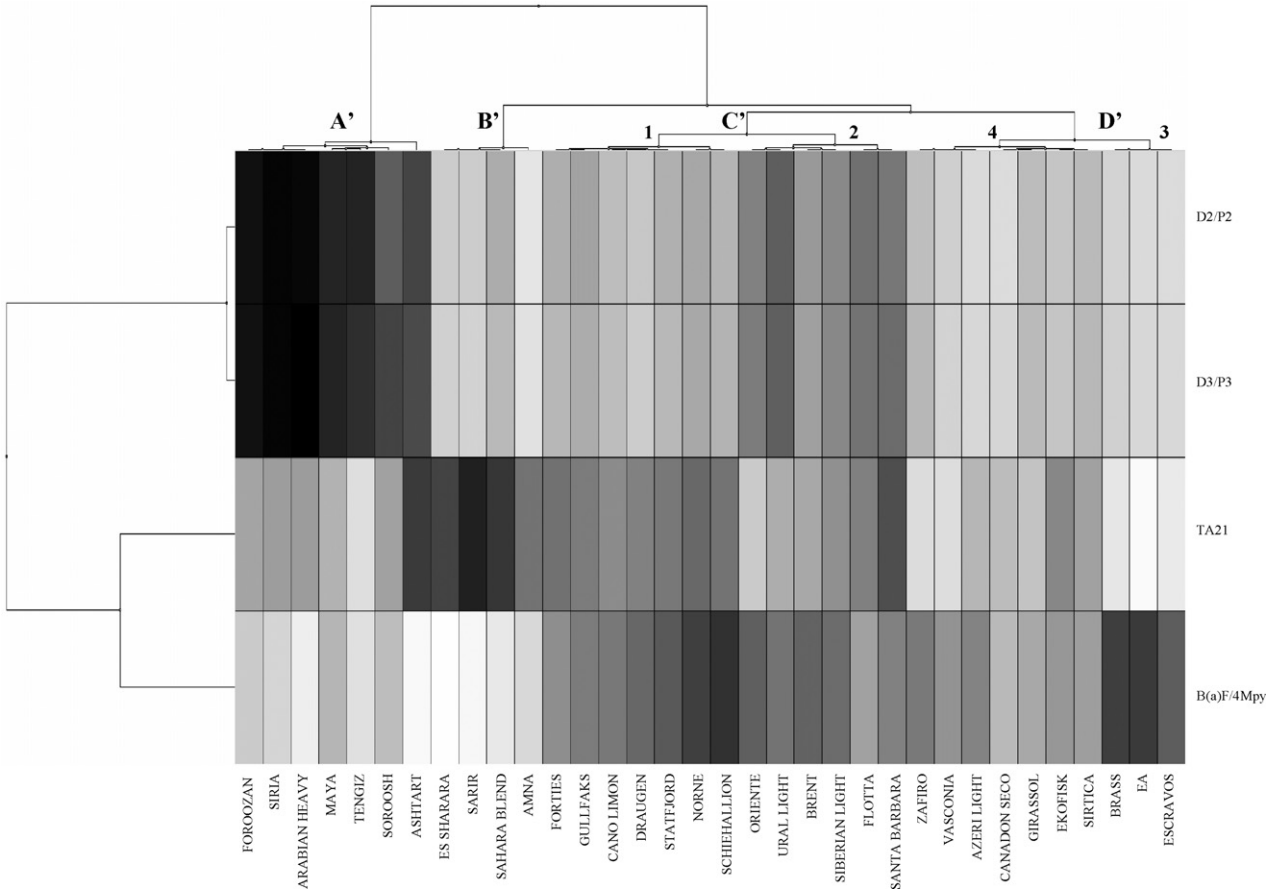

Fig. 7. Heatmap hierarchical clustering for the crude oils considering only the four selected variables. The experimental values of the DRs were coded with tiles whose shadow range from black (■, highest values) to white (□, lowest values).

inary trials. Note that in the following discussions the clusters obtained with the reduced set of variables have been denoted with a prime (′).

Cluster A′ matched with cluster A obtained using all variables, but for Ashtart (Tunissia) that was situated in a borderline position. This group was characterized by highest values on D2/P2 and D3/P3. Cluster B′ agrees with group B (considering all variables) although without the Amna crude oil (Lybia), nevertheless, this assignment is acceptable because Lybia is located North Africa, as for the other oils in group B′. This group got defined by highest values on TA21 and, at the same time, lowest values on B(a)F/4-Mpy.

Cluster C1′ coincided with subgroup C1 constituted by oils from the North Sea (also corresponding to group 5 of PCA) except for Caño Limón. This group was characterized by medium–high values

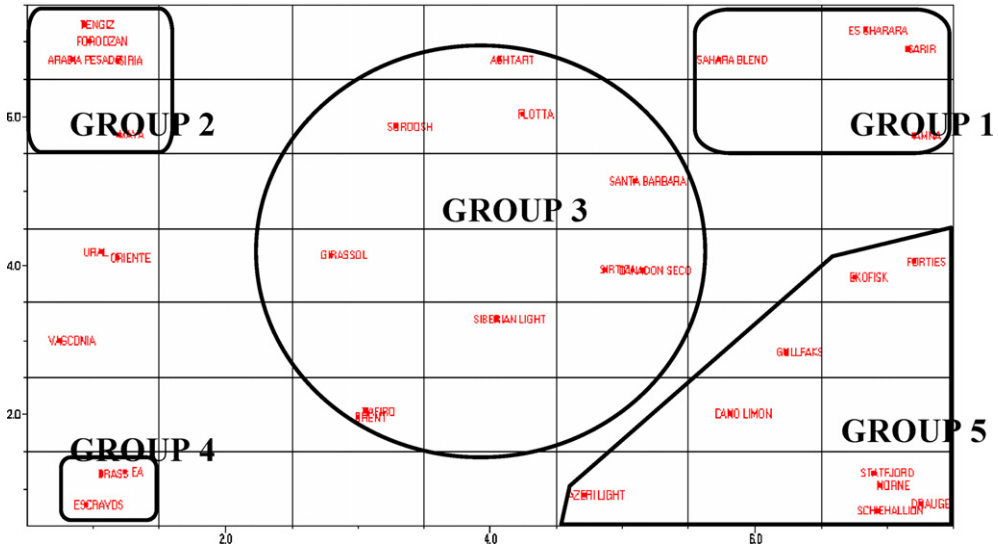

Fig. 8. SOM map for the crude oils dataset considering only the four selected variables.

on TA21 and B(a)F/4-Mpy and medium–low values on D2/P2 and D3/P3. Cluster C2' matched subgroup C2, which was defined by medium–high values on B(a)F/4-Mpy, D2/P2 and D3/P3.

Cluster D3′ agreed with subgroup D3 formed by Nigerian crude oils (group 4 from PCA). Samples giving rise to subgroup D3′ had highest values on B(a)F/4-Mpy and lowest on TA21. Cluster D4′ matched with subgroup D4, but for Amna (located close to cluster A1′) and Caño Limón (subgroup C1′). Subgroup D4′ presented medium values on B(a)F/4-Mpy and medium-high on D2/P2 and D3/P3.

Development of a Kohonen SOM using the 4 selected variables yielded a similar distribution of the samples as the all-variables-SOM, but for a rotation in the 2D map, which is not relevant (compare Figs. 5 and 8). Groups labelled as '1' were similar among them, but for Amna. Groups labelled as '2' were almost equal, except for Soroosh that was included at group '3' when only the 4 selected variables were considered. Groups termed as '3' became dissimilar because the Brent, Sirtica, Soroosh and Zafiro oils became included on group 3 derived from the selected variables, whereas the Azeri Light, Caño Limón, Oriente, Ural Light and Vasconia ones were not within it. The first two oils were included at group 5 (selected variables). The latter three oils were not included in any group. Groups labelled as '4' were almost similar but for the Zafiro oil. Groups '5' were similar but for the inclusion of the two oils mentioned above (Azeri Light and Caño Limón).

4. Conclusions

The use of univariate tools to analyse the suite of 28 diagnostic ratios included in this study was not successful in differentiating among the major oil production areas worldwide. Only some fuzzy trends could be appreciated in a classical ternary diagram. A PCA developed on the overall data set revealed four main groups of samples, linked to quite clear origins (North Sea, Nigeria, North Africa and Middle East), and another one without a clear geographical assignment. Similar groups were obtained by Kohonen neural networks and hierarchical cluster analysis (although the latter showed some disagreements with the PCA, likely caused by the different nature of the techniques and the fact that they do not consider the correlation between the variables).

An objective variable selection procedure based on Procrustes Rotation was undertaken to select a reduced set of DRs that comprised for most of the information in the original data without loosing relevant information. A reduced set of four DRs; namely; TA21, D2/P2, D3/P3 and B(a)F/4-Mpy characterized the crude oils and the groups they formed. The first ratio implied two triaromatic steroid biomarkers, whereas the last three involved ratios between PAHs. It is worth noting that the last diagnostic ratio has been included recently in the new technical report for identification of oil spills suited by the CEN European Committee.

The sample groups obtained by PCA, hierarchical clustering and grouping by Kohonen artificial neural networks were almost equal to those observed using all variables. Besides, some of the DRs objectively selected here had been mentioned in the literature as highly relevant to differentiate crude oils from different origins, which pointed out the utility of the reduced set of ratios to foresee the worldwide production area a crude oil came from.

Acknowledgements

The Spanish and Galician Governments (CIT-310200-2005-37 and 07MDS031103PR grants, respectively) are acknowledged by their support.

References

[1] P. Sun, M. Bao, G. Li, X. Wang, Y. Zhao, Q. Zhou, L. Cao, J. Chromatogr. A 1216 (2009) 830.
[2] K.E. Peters, J.M. Moldowan, The Biomarker Guide: Interpreting Molecular Fossils in Petroleum and Ancient Sediments, Prentice Hall, New Jersey, 1993.
[3] Z. Wang, M.F. Fingas, Mar. Pollut. Bull. 47 (2003) 423.
[4] Z. Wang, S.A. Stout, M. Fingas, Environ. Forensics 7 (2006) 105–146.
[5] K.E. Peters, C.C. Walters, J.M. Moldowan, The Biomarker Guide (volumes 1 and 2), Cambridge University Press, 2005.
[6] 'Biomarker focus' flyer, www.chiron.no, June 2010.
[7] A.O. Barakat, A.R. Mostafa, J. Rullkötter, A.R. Hegazi, Mar. Pollut. Bull. 38 (1999) 535.
[8] P.S. Daling, L.G. Faksness, A.B. Hansen, S.A. Stout, Environ. Forensics 3 (2002) 263.
[9] S.A. Stout, A.D. Uhler, K.J. McCarthy, Environ. Forensics 2 (2001) 87.
[10] J.H. Christensen, G. Tomasi, J. Chromatogr. A 1169 (2007) 1.
[11] R.C. Prince, D.L. Elmendorf, J.R. Lute, C.S. Hsu, C.E. Halth, J.D. Senius, G.J. Dechert, G.S. Douglas, E.L. Butler, Environ. Sci. Technol. 28 (1994) 142.
[12] T.J. Boyd, C.L. Osburn, K.J. Johnson, K.B. Birgl, R.B. Coffin, Environ. Sci. Technol. 40 (2006) 1916.
[13] G.S. Douglas, A.E. Bence, R.C. Prince, S.J. McMillen, E.L. Butler, Environ. Sci. Technol. 30 (1996) 2332.
[14] K.E. Peters, G.L. Scheuerman, C.Y. Lee, J.M. Moldowan, R.N. Reynolds, M.M. Pena, Energy Fuels 6 (1992) 560.
[15] Z.D. Wang, M. Fingas, M. Landriault, L. Sigouin, S. Grenon, Z. Zhang, Environ. Technol. 20 (1999) 851.
[16] Z.D. Wang, M. Fingas, L. Sigouin, Environ. Forensics 3 (2002) 251.
[17] T.C. Sauer, A.D. Uhler, Remediation 4 (1994) 431.
[18] K. Urdal, N.B. Vogt, S.P. Sporstol, R.G. Lichtenthaler, H. Mostad, K. Kolset, S. Nordenson, K. Esbensen, Mar. Pollut. Bull. 17 (1986) 366.
[19] ASTM D2709 – 96, Standard Test Method for Water and Sediment in Middle Distillate Fuels by Centrifuge, Annual Book of ASTM International Standards, 2006.
[20] L.G. Faksness, P.S. Daling, A.B. Hansen, CEN/BT/TF 120 Oil Spill Identification Summary Report: Round Robin Test Series B, SINTEF Report STF66 A02038, 2002.
[21] P.D. Boehm, G.S. Douglas, W.A. Burns, P.J. Mankiewicz, D.S. Page, A.E. Bence, Mar. Pollut. Bull. 34 (1997) 599.
[22] S.A. Stout, A.D. Uhler, T.G. Naymik, K.J. McCarthy, Environ. Sci. Technol. 32 (1998) 260A.
[23] CEN/TR 15522-2, Oil Spill Identification, Waterborne Petroleum and Petroleum Products, Part 2: Analytical Methodology and Interpretation of Results, Technical Report, 2006.
[24] L. Wilkinson, M. Friendly, The American Statistician 63 (2) (2009) 179–184.
[25] T. Kohonen, Self-Organizing Maps, Springer, Berlin, 2001.
[26] A.M. Fonseca, J.L. Biscaya, J. Aires-de-Sousa, A.M. Lobo, Anal. Chim. Acta 556 (2006) 374–382.
[27] W.J. Krzanowski, Principles of Multivariate Analysis; A User's Perspective, Revised ed., Clarendon Press, Oxford, England, 2000.
[28] J.M. Andrade, M.P. Gómez-Carracedo, W. Krzanowski, M. Kubista, Chemom. Intell. Lab. Syst. 72 (2004) 123.
[29] J.M. Deane, J.H. MacFie, J. Chemom. 3 (1989) 477.
[30] J.M. Andrade, D. Prada, S. Muniategui, B. Gomez, M. Pan, J. Chemom. 7 (1993) 427.
[31] R. Elias, A. Vieth, A. Riva, B. Horsfield, H. Wilkes, Org. Geochem. 38 (2007) 2111.
[32] W.-Y. Huang, W.G. Meinschein, Geochim. Cosmochim. Acta 43 (1979) 739.
[33] J.M. Moldowan, W.K. Seifert, E.J. Gallegos, AAPG Bull. 69 (1985) 1255.
[34] G. Dahlmann, Characteristic features of different oil types in oil spill identification, Berichte des BSH 31 (2003) 1–48.