Input Characterization of Sedimentary Organic Contaminants and Molecular Markers in the Northwestern Mediterranean Sea by Exploratory Data Analysis

JAUME S. I SALAU,[†] ROMÀ TAULER,*[,‡] JOSEP M. BAYONA,[†] AND IMMA TOLOSA[§]

Department of Environmental Chemistry, Centre d'Investigació i Desenvolupament (C.S.I.C), Jordi Girona 18, E-08034 Barcelona, Spain, Department of Analytical Chemistry, University of Barcelona, Diagonal 625, Barcelona E-08034, Spain, and IAEA, Marine Environmental Laboratory, B.P. 800, MC-98012, Monaco

Deposition zones of the NW Mediterranean were characterized according to the source of organic pollutants (i.e., UCM, PAHs, PCBs, DDTs) and lipidic compounds (i.e., alkanes and sterols) identified in surface sediments (31 samples) by principal component analysis (PCA) and hierarchical cluster analysis (HCA). Score plots of the two main principal components showed a cluster comprising the off-shore Barcelona and Rhône prodelta samples corresponding to the most polluted samples, while the remaining samples were clustered together. Loading plots revealed that most of the compounds were present in the first component except benzo[*ghi*]fluoranthene, the major DDT metabolites (i.e., DDE and DDD), and perylene, which was probably of diagenetic origin. In order to define further the cluster containing the most samples, a second data base that excluded the Rhône and offshore Barcelona samples was constructed. Score plot of the two principal components showed that three different depositional environments could be clearly defined, namely the Gulf of Lions, the Ebro prodelta, and the deep sea basin. Similar clustering was confirmed by HCA. The loading plots enabled riverine-transported compounds such as *n*-alkanes, PCBs, DDTs, sterols, and perylene (first component) to be distinguished from pyrolytic PAHs (second component). Furthermore, in order to obtain an apportionment of the inputs received to each station, a recently developed factor analysis multivariate curve resolution (MCR) method based on the alternating least squares (ALS) positive factorization of a data matrix was carried out for the first time on marine sediment samples. The ALS positive matrix factorization method enabled the apportionment of the environmental source of the main components of the compounds in the area of study.

Introduction

Traditionally, the molecular marker approach has been applied to source recognition of organic matter in the marine environment, which supposes that molecular markers have an unambiguous origin. In many cases, however, the lack of source specificity of many lipidic compounds (i.e., sterols, fatty acids) (*1*, *2*) and organic contaminants (PAHs) (*3*) has been found, which limits the use of such an approach in the source recognition of organic matter in the marine environment.

The usefulness of multivariate statistical techniques in source recognition in many environmental studies dealing with large data sets is known; but they are rarely applied. Zitko (*4*) illustrated from published data the potential of principal components analysis (PCA) application for expanding the interpretation of environmental data. Usually, either PCA or SIMCA (soft independent modeling of class analogy) have been successfully applied to source reconciliation of PAHs (*5*, *6*), polychlorinated dibenzo-*p*-dioxins and dibenzofurans (*7*−*10*) and lipids (*11*) from sediments. Other studies have applied PCA and factor analysis to a broad variety of chemical markers identified in several compartments for classification according to their origin (*12*). A crucial aspect for the source recognition from the organic patterns is the preservation of its integrity in the environment. This issue has been highlighted by Naes and Oug (*13*) for sedimentary PAHs. They concluded that compound-specific transformation reactions occurring during transport and incorporation into sediments contributed little to the total variance and did not suppress the source specific signal. Similarly, volatile hydrocarbons from vehicle emissions (*14*) and biogenic markers (*15*) were also useful as source discriminates because of their relatively high environmental stability.

Recently, Hopke (*16*) pointed out that source identification and quantitative mass apportioning of airborne particulate matter (commonly called receptor modeling) can be considered analogous to the spectrochemical mixture and multivariate calibration problems and that similar chemometrical tools could be brought to bear on the two types of problems. The present paper shows how a multivariate curve resolution method recently developed to solve mixture resolution problems (*17*, *18*) can also be used to derive source composition and source contribution profiles from environmental multicomponent analysis. The method is based on alternating least squares positive matrix factorization (*19*) and source profiles optimization to fit experimental data.

The objectives of this study were 3-fold: (i) the classification of the different deposition zones of study (NW Mediterranean Sea) according to the source of organic compounds occurring in surface sediments, (ii) classification of a large variety of organic compounds according to their different origins, and (iii) apportionment of input sources in the different samples. For this purpose, a data set containing concentrations of biogenic (i.e., odd carbon-numbered *n*-alkanes, sterols) and individual anthropogenic compounds (i.e., even carbon-numbered *n*-alkanes, PCBs, PAHs, DDTs, etc.) from surface sediments collected from the NW Mediterranean Sea was used in this study. Consequently, an exploratory study by PCA, followed by hierarchical cluster analyses (HCA) and alternating least squares (ALS) positive matrix factorization, was carried out.

Experimental Section

Data Origin. The data set selected for this study was taken from Tolosa *et al.* (*20*, *21*) and Lipiatou and Saliot (*22*, *23*). The sample grid covers the NW Mediterranean basin focusing on the Ebro and Rhône river prodeltas, which are the major sources of land-based organic matter in the region. Figure 1 shows the sampling site location. A detailed description of the sampling protocol, sample handling, analyses, concen-

* Corresponding author fax: 34-3-402.12.33; e-mail: roma@quimio.qui.ub.es.

[†] Centre d'Investigació i Desenvolupament.

[‡] University of Barcelona.

[§] IAEA.

S0013-936X(97)00231-9 CCC: $14.00

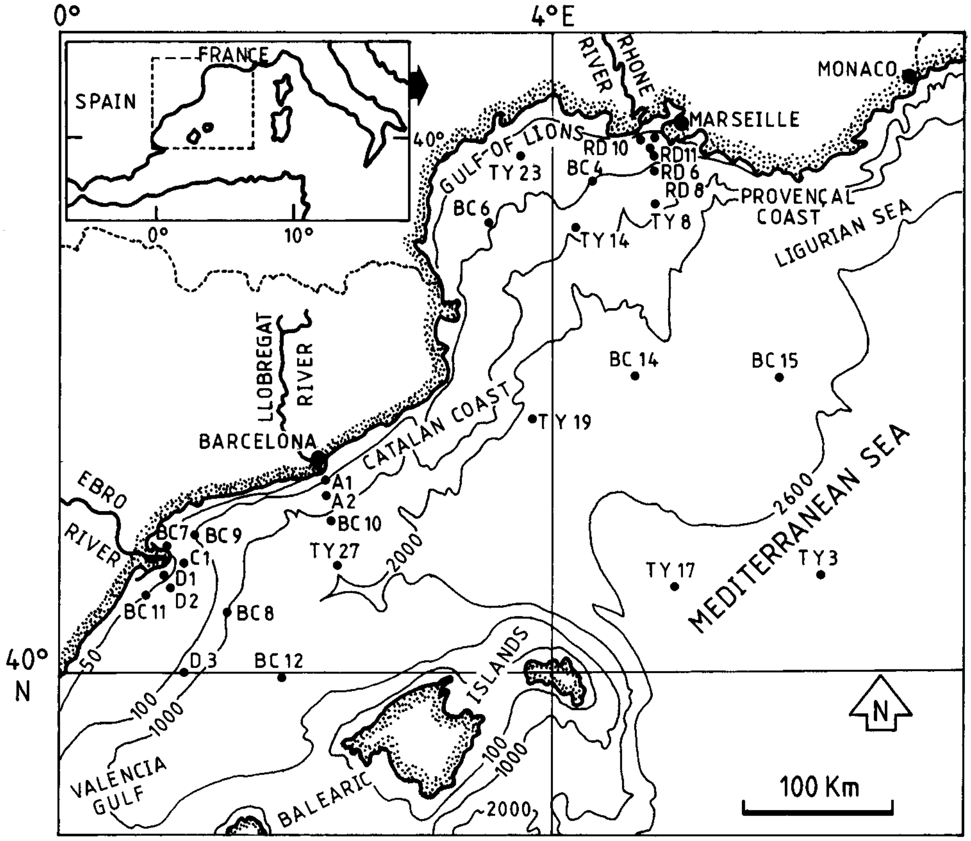

FIGURE 1. Area of study and sampling site locations.

trations, fluxes, and loads of pollutants in this region can be found elsewhere (*20*−*23*). The initial exploratory data analysis carried out covered the entire geographical area: a data set containing 31 samples and 59 compounds (*n*-alkanes, PCBs, PAHs, chlorinated pesticides) (1829 values) was considered. A second data set was constructed, excluding the Barcelona and Rhône prodelta samples and including 22 samples and 96 compounds (*n*-alkanes, PCBs, PAHs, sulfur-containing PAHs, chlorinated pesticides, sterols) (2112 values) (Table 1).

Methodology. Data treatment and analysis involved four steps: (i) data screening and pretreatment; (ii) explorative data description by PCA, (iii) HCA, and (iv) ALS positive matrix factorization. The Pirouette (*24*) and home-made MATLAB (*25*) data analysis software were extensively used for most of the calculations.

(i) Data Screening and Data Pretreatment. A data matrix, with as many rows as samples analyzed and as many columns as variables or analyte concentrations measured, was built up. The problem of unsuitable data sets has been addressed in the literature by preliminary selection of the samples and compounds to be included in the calculations using the already available geochemical information. Data pretreatment was different for the PCA and ALS positive matrix factorization. While in PCA, the data were autoscaled, i.e., standardized (scaled to the same units dividing each element of the matrix by the standard deviation of its column) and mean centered (each element subtracted by its mean column); in the ALS treatment, data were standardized but not mean centered. Therefore, in the ALS treatment of data, information about origin (zero point of the data scale) is not lost and apportionment (*16, 26, 27*) is possible.

(ii) Data Exploration and Description by Principal Component Analysis (PCA). PCA is used as an explorative tool to investigate how many components are needed to explain variance in observed data. The number of components found by pure mathematical means is then related to the number of real sources of data variation. Special attention is paid to those principal components that explain the larger parts of data variance. Components explaining little data variance (i.e., less than 5%) are not investigated and assumed to be mostly due to background and noise contributions. Extremely useful tools for data exploration are the score and loading plots derived directly from PCA analysis that, respectively, map samples and variables in the new vector space defined by the principal components. Score plots allow sample identification, checking whether they are typical or outliers, similar or dissimilar. Score plots also enable sample clusters (groupings) to be searched for. From loading plots, the more important variables can be identified. Variables with large loadings, close to each other, and along the same straight line through the origin covary; if they are on the same side of the origin, they covary in a positive way, whereas if they lie on opposite sides, they are correlated negatively. Interpretation of clusters of samples in the score plot is simultaneously done by studying the corresponding loading plot.

(iii) Hierarchical Cluster Analysis (HCA). HCA was used as a confirmatory tool to model the groupings or clusters of experimental data previously identified by PCA. The primary purpose of HCA was to present the data so as to emphasize the natural groupings in the data set. The presentation of HCA analysis is usually performed in the form of a dendogram, making possible the visualization of clusters and correlation among samples. Clusters are defined through distances (Euclidean), differences, or similarities between two samples at each of the measured variables (concentration of the compounds). There are different ways to group the samples according to different distance measurements and methods of linking samples (*24*).

(iv) Alternating Least Squares (ALS) Positive Matrix Factorization. A different approach based on factor analysis (*28*) principles was proposed in the second data set. The method is based on a multivariate curve resolution (MCR) method recently proposed to solve mixture analysis problems in different fields (*17, 18*). Like in PCA and other traditional

TABLE 1. Compounds Analyzed and Code Identification[a]

first	second	compounds	mean	SD
1	1	*n*-C16	14.7	7, 70
2	2	*n*-C17	33.7	23, 15
3	3	*n*-C18	22.4	9, 07
4	4	*n*-C19	23.5	11, 07
5	5	*n*-C20	19.3	7, 79
6	6	*n*-C21	32.3	16, 72
7	7	*n*-C22	22.0	8, 17
8	8	*n*-C23	25.3	9, 39
9	9	*n*-C24	20.5	8, 31
10	10	*n*-C25	37.3	13, 85
11	11	*n*-C26	22.3	8, 74
12	12	*n*-C27	75.6	36, 06
13	13	*n*-C28	32.7	14, 42
14	14	*n*-C29	209.9	77, 87
15	15	*n*-C30	49.8	19, 77
16	16	*n*-C31	236.1	81, 47
17	17	*n*-C32	32.1	13, 92
18	18	*n*-C33	104.5	37, 71
19	19	*n*-C34	20.6	11, 04
20	20	*n*-C35	51.0	19, 07
21	21	*n*-C36	15.3	13, 97
22	22	*n*-C37	17.3	14, 12
23	23	*n*-C38	14.6	13, 03
24	24	*n*-C39	11.4	11, 76
	25	UCM (unresolved carbon mixture)	12324	4994, 65
	26	pristane	25.3	20, 58
	27	phytane	8.4	4, 91
	28	fluoranthene	0.5	0, 50
25	29	phenanthrene	18.9	12, 21
26	30	anthracene	1.8	1, 52
	31	methylphenanthrene	16.4	9, 89
	32	dimethylphenanthrenes	10.0	6, 84
27	33	fluoranthene	33.8	18, 32
	34	acephenantrylene	1.2	1, 13
28	35	pyrene	25.7	14, 00
	36	methylfluoranthenes	6.9	4, 18
	37	benzo[*a*]fluorene	4.6	3, 10
	38	benzo[*b*]fluorene	1.9	1, 33
29	39	retene	1.7	1, 25
30	40	benzo[*b*]phenanthrene	2.2	1, 34
31	41	benz[*a*]anthracene	15.0	9, 71
32	42	crysene + triphenylene	35.5	24, 51
33	43	benzo[*j*+*b*+*k*]fluoranthenes	80.7	56, 81
34	44	benzo[*a*]fluoranthene	4.6	3, 48
35	45	benzo[*e*]pyrene	27.7	17, 47
36	46	benzo[*a*]pyrene	18.0	13, 61
37	47	perylene	28.0	40, 06
	48	indeno[7,1,2,3-*cdef*]chrysene	4.9	4, 77
38	49	indeno[1,2,3,-*cd*]pyrene	29.7	20, 65
39	50	benzo[*ghi*]perylene	25.0	16, 44
40	51	benzo[*ghi*]fluoranthene	6.6	4, 38
	52	cyclopenta[*cd*]pyrene	0.6	0, 78
	53	dibenzoanthracenes	7.9	7, 55
	54	benzo[*b*]chrysene	1.2	1, 27
	55	coronene	11.5	7, 20
	56	302	23.8	16, 78
	57	naphtho[1,2,-*b*]thiophene	0.2	0, 25
41	58	dibenzothiophene	0.9	0, 82
	59	naphtho[2,1-*b*]thiophene	0.1	0, 10
	60	4-methyldibenzothiophene	0.7	0, 50
	61	3,2-methyldibenzothiophene	0.4	0, 31
	62	1-methyldibenzothiophene	0.2	0, 11
43	63	benzo[*b*]naphtho[2,1-*d*]thiophene	6.7	4, 84
	64	benzo[*b*]naphtho[1,2-*d*]thiophene	1.3	0, 80
44	65	benzo[*b*]naphtho[2,3-*b*]thiophene	1.2	0, 92
45	66	PCB-52	0.3	0, 40
46	67	PCB-101	0.5	0, 51
47	68	PCB-118	1.3	1, 60
48	69	PCB-153	1.1	1, 29
49	70	PCB-138	1.4	1, 52
50	71	PCB-187	0.7	0, 89
51	72	PCB-128	0.2	0, 20
52	73	PCB-180	1.3	1, 83
53	74	PCB-170	0.9	1, 32
54	75	*o*,*p*′-DDD	1.9	3, 74

TABLE 1 (Continued)

first	second	compounds	mean	SD
55	76	*o,p′*-DDE	0.2	0, 20
56	77	*o,p′*-DDT	0.4	0, 78
57	78	*p,p′*-DDE	2.0	1, 78
58	79	*p,p′*-DDD	5.1	10, 44
59	80	*p,p′*-DDT	8.1	13, 67
	81	hexaclorobenzene	2.2	4, 41
	82	hexaclorohexane	0.0	0, 10
	83	lindane	0.1	0, 30
	84	octachloroestyrene	0.2	0, 29
	85	27-nor-24-methylcholesta-5α,22(*E*)-dien-3*β*-ol	90.3	73, 49
	86	cholesta-5α,22(*E*)-dien-3*β*-ol	194.5	169, 63
	87	cholesterol	462.3	506, 97
	88	cholestanol	206.2	160, 40
	89	brassicasterol	359.5	365, 99
	90	24-methyl-5α(*H*)-cholest-22(*E*)-en-3*β*-ol	130.2	128, 24
	91	24-methylhcolest-5-en-3*β*-ol	95.9	70, 15
	92	stigmasterol	269.9	206, 31
	93	24-ethyl-5α-cholest-22-en-3*β*-ol	66.9	54, 91
	94	*β*-sitosterol	499.0	395, 06
	95	24-ethyl-5α-cholestan-3*β*-ol	209.4	189, 79
	96	dinosterol	172.6	173, 52

[a] First and second columns correspond to the first and second data set, respectively. Mean (ng/g) and standard deviation (SD) correspond to the second data set, without Barcelona and Rhône area samples.

factor analysis (FA) methods, in MCR methods, the experimental data are arranged in a single data matrix. No regression is attempted between two or more different data blocks, like in multilinear regression (MLR), principal components regression (PCR), partial least squares regression (PLSR), or other multivariate regression methods based or not in matrix factor decomposition. The original data matrix **D** consisting of NS samples (NS = 22, rows) and NV variables (NV = 96, columns) (compounds analyzed) is decomposed as a product of two factor matrices, **R** and **C**, related respectively to the rows and the columns of the original data matrix **D**, where **E** is the residual data matrix containing unexplained data variance:

$$\mathbf{D}_{(\mathrm{NS,NV})} = \mathbf{R}_{(\mathrm{NS,NC})}\,\mathbf{C}_{(\mathrm{NC,NV})} + \mathbf{E}_{(\mathrm{NS,NV})} \qquad (1)$$

This equation implies that each data element d_{ij} in **D** is written as a sum of $k = 1, ..., \mathrm{NC}$ linear contributions:

$$d_{ij} = \sum_{k=1}^{\mathrm{NC}} r_{ik}c_{kj} + e_{ik} \qquad (2)$$

In the particular case under study, these NC linear contributions are supposed to be the different environmental sources of the chemical components analyzed in the different samples. Each source (k) is characterized by a profile of the concentrations of the analyzed compounds j (c_{kj}), i.e., by a linear combination of the measured variables; these profiles will be called *source composition profiles*. And each source (k) is also characterized by a distribution profile of relative contributions among samples, or which is the same, each source k contributes a certain amount to the total input content of each sample i (r_{ik}); these profiles will be called *source contribution profiles*. **E** accounts for the unexplained variance when using a preselected number of sources in **R** and **C** for optimal description and reproduction of experimental data matrix **D**, respectively. In fact **E** has apart from experimental noise all the background contributions not explained by the NC sources. In this type of environmental study, in contrast to spectrometric mixture analysis studies (*18*), **E** still has a significant amount of unexplained data variance from background source contributions that cannot easily be identified when using a small number of contributions. Solving eq 1 for **R** and **C** for a known **D** matrix is ambiguous (factor analysis ambiguity, see refs 17, 18, 29, and 30), since there are an infinite number of possible **R** and **C** matrices which, when multiplied, reproduce the same data matrix **D**. However, if a set of natural physical constraints is imposed, the number of possible solutions for contribution and composition profiles, **R** and **C** matrices, is drastically reduced. A first constraint is to find only those solutions that minimize the unexplained variance **E** in eq 1, for instance, using a least squares criteria. A second constraint is that the predicted source contributions to each data sample (source contribution profiles) must be non-negative (i.e., $r_{ik} > 0$). And the third constraint is that concentrations of the compounds in the source composition profiles must also be positive (i.e., $c_{ki} > 0$). These constraints are similar to those usually imposed in mixture analysis (*17, 18, 29, 30*) and to those recently proposed in environmental and source apportionment studies like positive matrix factorization (*16, 19, 26, 27*).

The particular implementation of these constraints in the proposed alternating least squares procedure is summarized in the following steps:

(1) Experimental data were arranged in a matrix (**D**) with as many rows as samples analyzed (NS) and as many columns as variables or analyte concentrations measured (NV). The columns of the original data matrix (concentrations of a particular component in the different samples) were scaled to the same units by dividing each element of the matrix by the standard deviation of its column. The new data matrix was not centered in order not to miss the reference information about the data center (as in apportionment studies; *16, 19, 26*).

(2) The number of components (environmental sources) are initially estimated as in PCA and then from the visual inspection of the size of eigenvalues of the scaled matrix. For a set of NC components, the reproduced data matrix **D*** is calculated from the factor decomposition of the experimental matrix **D**

$$\mathbf{D} = \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{E}_{\mathrm{PCA}} = \mathbf{D}^{*} + \mathbf{E}_{\mathrm{PCA}} \qquad (3)$$

where **U** is the matrix of scores and $\mathbf{V}^{\mathrm{T}}$ is the matrix of loadings and $\mathbf{E}_{\mathrm{PCA}}$ is the residuals matrix containing the unexplained variance for the selected number of components (NC).

The initial assumption is that the estimated number of components (NC) is equal to the number of independent and more significant environmental sources of the analyzed compounds; i.e., N_{PCA} = NC.

(3) Since ALS is an iterative method, it has to be provided with starting values. Like in other iterative optimization methods, the proposed ALS method encounters to a certain extent the problems of iterative estimation: local minima, convergence problems, slowness in the iterations. In this context, the selection of starting values is important. An initial estimation of the composition profiles of each environmental source ($\mathbf{C}^{inic}$) is taken directly from the original data matrix, looking for those samples that are the 'purest' representation of the composition of the sources. This search is carried out using a similar approach to that proposed in the SIMPLISMA method (*31, 32*). These NS purest samples detected in the data matrix provide initial estimations of the NS purest composition profiles of the sources sought ($\mathbf{C}^{inic} = \mathbf{C}$ in eq 1). These initial estimations are not however optimal estimations from a least squares sense (minimum variance), and they can be improved considerably using the proposed ALS optimization method described in the next section.

(4) An ALS optimization is started using the two equations:

$$\mathbf{R} = \mathbf{D}^{*}\mathbf{C}^{+} \tag{4}$$

and

$$\mathbf{C} = \mathbf{R}^{+}\mathbf{D}^{*} \tag{5}$$

where $\mathbf{C}^{+}$ and $\mathbf{R}^{+}$ are the pseudoinverse (*33*) matrices of **C** and **R** respectively. Note that $\mathbf{D}^{*}$ (eq 3) is used instead of **D** to improve stability during the calculations. At each iteration of the optimization, the two non-negativity constraints are applied:

$$\mathbf{R} > 0$$

and

$$\mathbf{C} > 0$$

The optimization is carried out iteratively until no further improvement is found in the data fitting and convergence is achieved. Ideally, the ALS solutions and the PCA solutions for **R** and **C** will give a similar fit to the original experimental data matrix, i.e., the data matrix reproduced by the positively constrained ALS optimization procedure ($\mathbf{D}_{ALS}$) will be very close to that reproduced by the PCA method ($\mathbf{D}^{*}$). However, the PCA solution will be different from the ALS solution, since the PCA solutions are constrained to be orthogonal whereas ALS solutions are constrained to be positive but not orthogonal. The comparison and interpretation of the results with both methods will be useful.

Finally, the matrix $\mathbf{E} = \mathbf{D} - \mathbf{D}_{ALS} = \mathbf{D} - \mathbf{RC}$ has the unexplained data variance using the ALS model. This remaining data variance cannot be explained significantly by a reduced set of linear environmental sources and is assumed to be caused by a large number of small background environmental source contributions that cannot be individually distinguished from noise.

The unambiguous characterization of contribution (**R** matrix) and composition (**C** matrix) profiles from the analysis of a single data matrix (**D** matrix) using MCR methods assumes that (a) the sources are identifiable; (b) they contribute to the data variance linearly, i.e., **R** and **C** matrices give information about the contributions and composition of the sources; and (c) the decomposition of the matrix **D** in **R** and **C** factor matrices is achieved correctly by the proposed ALS method. The last point deserves further discussion. Decomposition of a single data matrix by FA methods gives two types of ambiguities: rotational and intensity ambiguities. Rotational ambiguities mean that the recovered solutions are a linear combination of the true ones. Intensity ambiguities mean that the recovered solutions are scale undetermined. Rotational ambiguities can be solved under special conditions related with the selectivity of the measurements (*17*). Also the application of natural constraints like non-negativity reduce considerably the number of possible solutions of eqs 4 and 5, especially for real data. Intensity ambiguities cannot be solved in the analysis of a single data matrix unless external information about the scale of the measurements is provided. The results obtained at the end of the optimization process by ALS must be checked for lack of fit (unexplained variance for the selected number of components), for different starting values giving the same results, and for interpretability. If all these conditions are solved satisfactorily, the recovered solutions will be very close to the true physically underlying phenomena under study and eventually give the true ones (*17, 18*).

TABLE 2. PCA Analysis for First Data Set[a]

	%	cum		%	cum
PC1	75.4	75.4	PC3	5.2	89.4
PC2	8.8	84.2	PC4	3.8	93.2

[a] %, percent of variance; cum, cumulative variance.

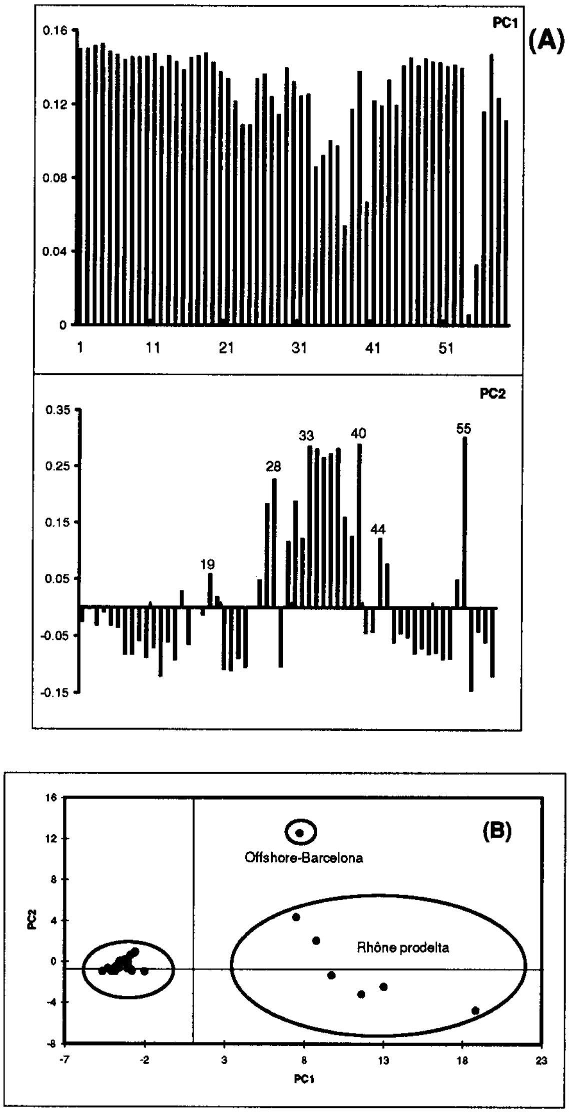

FIGURE 2. (A) Loadings and (B) scores plots of the PCA analysis of the first data set (all the analyzed samples). For variable code identification in the loading plots, see first column of Table 1.

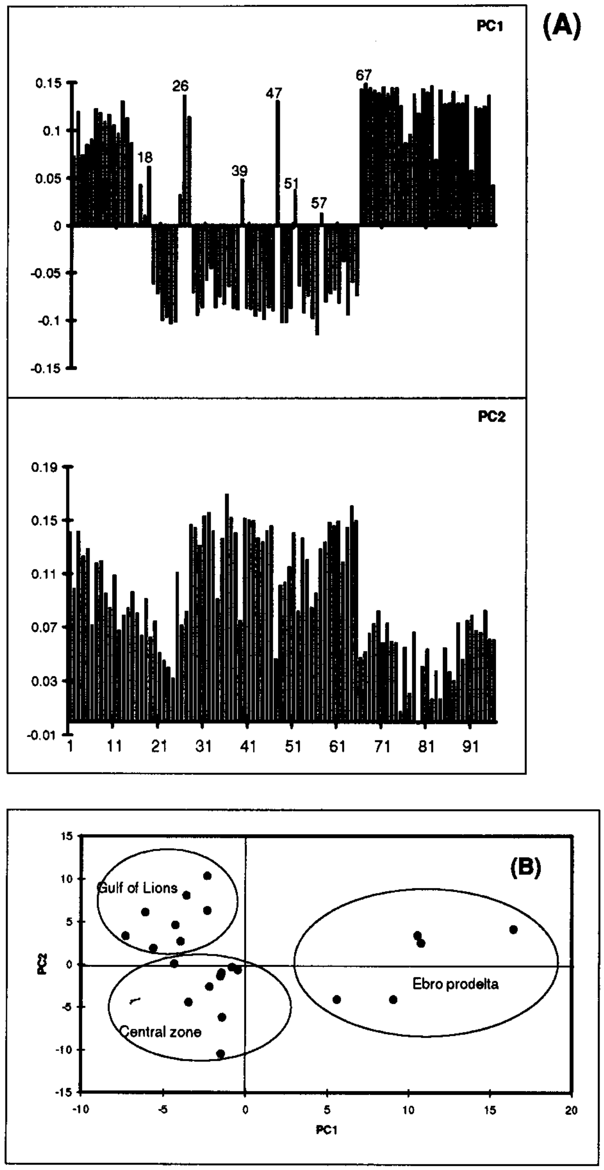

FIGURE 3. (A) Loadings and (B) scores plots of the PCA analysis of the second data set (samples from Barcelona and Rhône area excluded). For variable code identification in the loading plots, see second column of Table 1.

Results and Discussion

Principal Component Analysis (PCA). An initial exploratory data analysis, covering the entire geographical area and considering the following individual molecular markers: *n*-alkanes, PAHs, PCBs and DDTs, was carried out. The first two principal components (PCs) account for 84.2% of the total variance (Table 2). The combination of variables and their loadings are shown in Figure 2A. Almost all the variables except *o,p*′-DDD, *o,p*′-DDE, benzo[*ghi*]fluoranthene, and perylene contributed to the first PC, which accounts for 75.4% of the data variance, and are positively correlated. These results indicate that all pollution inputs were evenly distributed in the area of study (i.e., urban and industrial), probably due to their geographical proximity (Rhône River, Gulf of Lions, and Barcelona urban area, Figure 1). Furthermore, the lower contribution of *o,p*′-DDD and *o,p*′-DDE to the first variable could be attributable to the metabolic origin in *o,p*′-DDT of these compounds (*34, 35*). Similarly perylene, which has a dual—either diagenetic or pyrolitic—origin (*36*), has a smaller contribution to the first PC than the remaining pyrolytic PAHs. Benzo[*ghi*]fluoranthene also contributes little to the first PC, which suggests that this compound has a specific environmental pathway. Benzonaphthothiophenes (**43** and **44** in Table 1) are positively correlated with other pyrolytic PAHs (e.g., fluoranthene, **27**; pyrene, **28**; benzofluoranthenes, **33** and **34**; benzopyrenes, **35** and **36**; perylene, **37**; and benzo[*ghi*]perylene, **39**). Thus it could indicate that benzonaphthothiophenes exhibited a common pyrolytic source with other pyrolytic PAHs. Furthermore, dibenzothiophene (**41**) is correlated with fossil hydrocarbons such as *n*-alkanes and phenantrene, which might indicate a common origin. The correlation between fossil hydrocarbons with DDTs (compounds **55**−**59**) and PCBs (compounds **45**−**53**) is an evidence of a predominant riverine transport and continental runoff of these compounds in the area of study.

TABLE 3. PCA Analysis and ALS Analysis for Second Data Set[a]

	PCA (autoscaled data)[b]		PCA (standardized data)[c]		ALS (standardized data)	
	%	cum	%	cum	%	cum
1st	41.4	41.4	53.6	53.6	19.1	19.15
2nd	26.8	68.2	14.9	68.5	27.0	46.19
3rd	8.4	76.6	4.8	73.3	26.8	73.03

[a] %, percent of variance; cum, cumulative variance. [b] PCA analysis is performed after mean centering and standardization as in Table 2 for data set 1. [c] PCA analysis is perfomed after standardization without mean centering to compare with ALS data analysis.

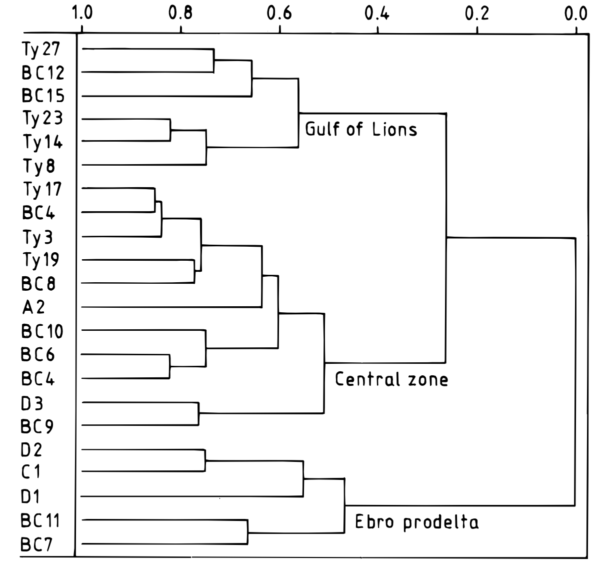

FIGURE 4. Dendogram obtained from the HCA of the second data set, without Barcelona and Rhône area samples.

The second PC accounts for 8.8% of the total variance, which is attributable to a positive contribution of some of the three−five aromatic ring PAHs (i.e., benzofluoranthene isomers, benzopyrenes, perylene, fluoranthene, pyrene, anthracene), mostly of pyrolytic origin, and to the transformation products of DDT (i.e., *o,p*′-DDE and *o,p*′-DDD) (Figure 2A). Parent pesticides are negatively correlated according to this PC (*o,p*′-DDT). Retene, usually associated either to wood combustion or diagenetic origin (*37*), is negatively correlated with pyrolytic PAHs.

By plotting the scores of the first two PCs, the areas of study can be grouped in three clusters (Figure 2B): (i) a first cluster with a single sample located off-shore from Barcelona that contained the highest amount of PC-2, (ii) a second cluster formed by the samples located in front of the Rhône prodelta characterized by the highest values of PC-1, and (iii) a dense third cluster containing the remaining samples that have the lowest values for both PCs. These results are consistent with (i) the higher pollution levels found off-shore Barcelona and

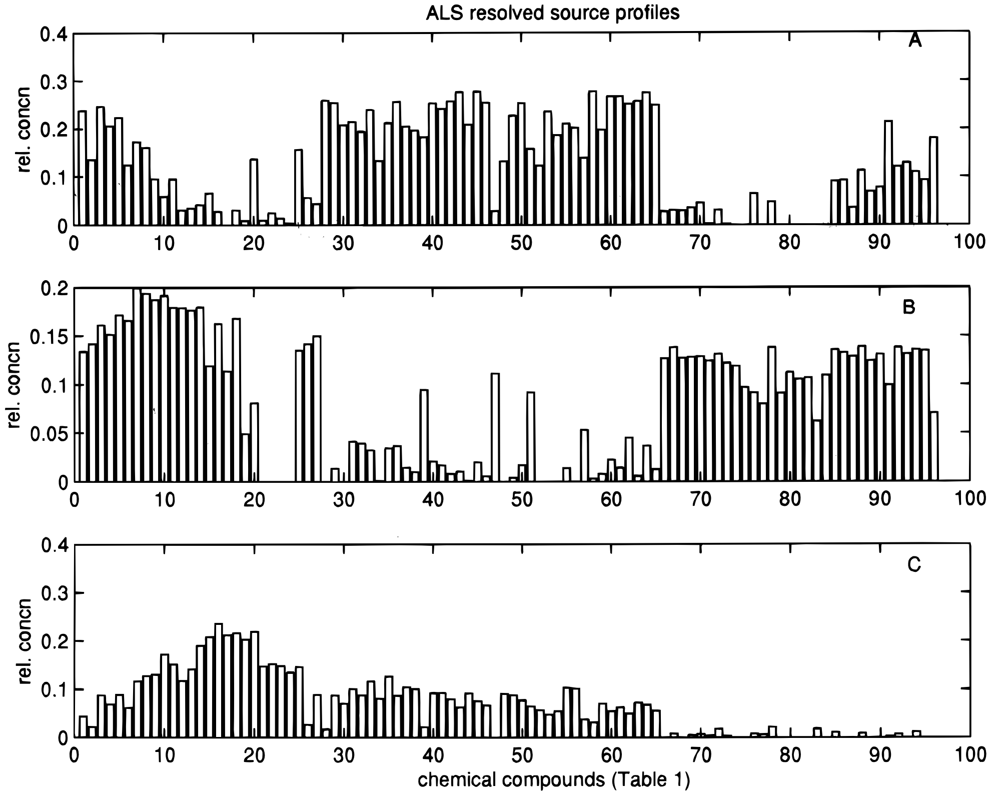

FIGURE 5. Resolved composition profiles (matrix C in eq 1 or c_k profiles in eq 2) of the three sources (A−C) by the ALS method. The numbers on the *X* axis are the variable code identification given in the second column of Table 1. The *Y* axis gives the relative concentrations in arbitrary units.

the Rhône prodelta and (ii) the pollution of these areas associated with different sources. Indeed, the higher contribution of PC-2 in the sample off-shore Barcelona is accounted for by the *o,p'*-DDE associated with local sources of pollution (*20*) and the pyrolytic PAHs that come from mobile sources (*21*).

In order to get further insight into the sources of the remaining samples, the off-shore Barcelona and Rhône prodelta samples were skipped from the data matrix. Thus, a new data subset containing the former variables and sterols was built up (second column in Table 1). The resultant new submatrix had 96 variables (compounds) and 22 samples. In this case, the first four PCs accounted for 80.5% of the total data variance. The positive contribution to the first PCA is accounted for now by biogenic or anthropogenic origin compounds coming from land-based sources (Figure 3A): (i) higher plants (n-C_{31}, n-C_{33}, and 24-ethylcholest-5-en-3β-ol), (ii) fossil sources of hydrocarbons (n-C_{17}−n-C_{29}, pristane, phytane, UCM), (iii) wood combustion (retene), (iv) diagenetic (perylene), (v) coal mining (1-naphthothiophene), (vi) industrial origin (PCBs), and (vii) pesticides (lindane, DDT, and its metabolites). The most probable route of transport of these compounds could be associated to continental runoff and/or river transport. A negatively correlated contribution was evident for the rest of the PAHs and higher molecular weight n-alkanes (n-C_{34-39}). While the former are associated with combustion processes and atmospheric transport, the latter can be associated to tanker ballast operations carried out in the open sea (*21*). This first PC represents 41.4% of the total variance (Table 3). The score plots in Figure 3B show that the Ebro prodelta samples displayed the highest contribution on the PC-1 value.

The second PC represents 26.8% of the total variance and shows the contribution of n-alkanes, pyrolytic PAHs, and to a lesser extent PCBs, DDTs, and sterols. The higher contribution of this PC in the Gulf of Lions samples is consistent with a predominant contribution of atmospheric deposition for this PC. In addition, samples located in the deep basin formed another cluster that is characterized by intermediate values of both PCs. This grouping is consistent with the contribution of both atmospheric and advective transport from the continental shelf.

Score plot PC1 *vs* PC2 clearly shows three clusters of samples identified as Ebro prodelta, Central Zone, and Gulf of Lions (Figure 3B). Loading plot PC1 *vs* PC2 shows two groups of contributions separated by PC1; one of which is mainly composed of n-alkanes and PAHs, and the other is composed of other n-alkanes, PCBs, and DDTs. Comparison of scores and loadings plots lead to the conclusion that Ebro prodelta samples are mostly distinguished by the contribution of n-alkanes, PCBs, DDTs, and pristane and that the distinction between samples from the Gulf of Lions and the Central Zone is due to differing amounts of n-alkanes and PAHs.

Hierarchical Cluster Analysis (HCA). HCA results are shown in Figure 4. The linking method that gave the most similar results to those obtained by PCA was the incremental link method (*24*). This method uses a sum of squares approach for calculating the nearest cluster. Other linking methods, like the single link method, the centroid link method, or the complete link method among others produced slightly different groups. Three sample groups could be modeled from data similarities in line with the three clusters found by PCA. Two samples (Ty-27 and BC-12 in Figure 1), which are located at the bottom of the slope of the Ebro prodelta, are grouped with the Gulf of Lions cluster. These two samples displayed high composition similarity with the remaining samples from the Gulf of Lions (Figure 3B). Although no definitive explanation can be given at this stage, these results may be explained by the fact that the slope of the Ebro prodelta zone is characterized by large sedimentation rates due to advective transport from eroded sediments off the slope and continental shelf (*38*). Furthermore, another sample collected on the edge of the Rhône prodelta showed greater similarity with HCA to the deep basin samples, instead of with the Gulf

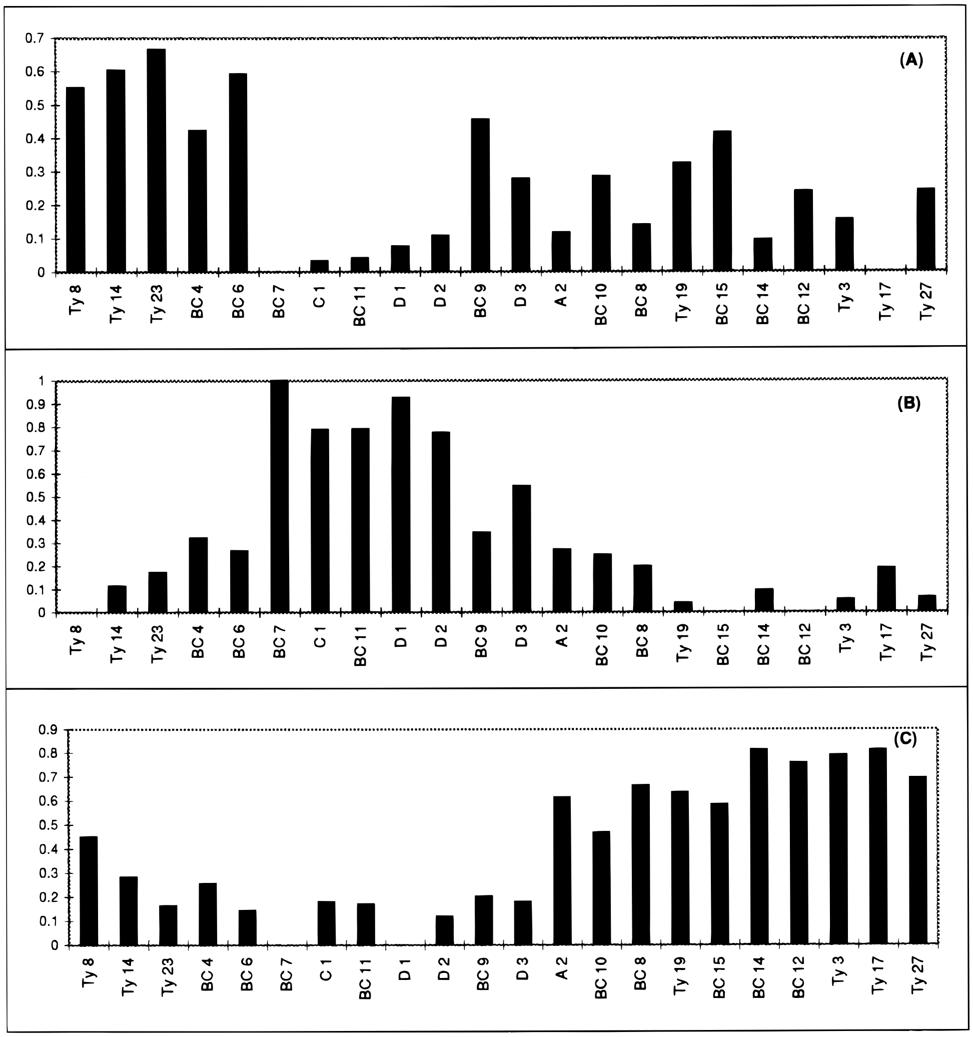

FIGURE 6. Resolved contribution profiles (matrix R in eq 1 or r_k profiles in eq 2) of the three sources A−C) according to the sampling site obtained by the ALS method. Identification of the sampling sites are given in the *X* axis (see coding in Figure 1).

of Lions ones (BC-4 in Figure 1). As the prevalent sea current is NW−SE (*39*), this sample could have been affected by the Rhône plume. Therefore, both river transport and atmospheric deposition could be the main routes of organic matter transport into this region. Consequently, its grouping is closely related to the deep basin samples where both inputs are apparent.

Alternating Least Squares (ALS) Positive Matrix Factorization. In the study of the reduced data matrix (without the Barcelona and Rhône samples and incorporating sterol variables, second column in Table 1), three main components were initially considered in accordance with previous results with PCA. The initial basic assumption was that these three contributions could be associated with three different specific environmental sources and that the remaining unexplained data variance (26.7% in Table 3) was caused by background contributions unconnected with significant environmental sources.

The initial estimations of the composition profiles of the three proposed environmental sources (**C** in eq 1) were made using the purest samples detected from the data matrix (*31, 32*). These three initial composition profiles corresponded to the chemical concentrations of samples BC7, D3, and BC15 shown in Figure 1. BC7 is a characteristic sample from the Ebro Delta zone, D3 is a sample from the slope, and BC15 is from the deep basin. The application of the MCR-ALS method previously described (*17, 18*), using these three initial estimations of the source composition profiles (see Methods section iv), gave an optimum set of profiles for each of these three identified environmental sources. The amount of experimental data variance finally explained by the ALS method was similar to that obtained with the PCA method (73% for non-centered standardized data) (Table 3), which confirms that about the same amount of data variance can be explained by both methods. However, the recovered profiles can be interpreted more easily with the proposed ALS method than with the PCA method, as is shown below.

In Figure 5, the three ALS resolved source composition profiles are given. The source profile A has high concentration inputs, mostly of PAHs (**26**−**46** in second column of Table 1), which are assumed to be of anthropogenic origin. Perylene (**47**), which has a dual origin—either diagenetic or pyrolytic (*36*), makes a small contribution to the first component. UCM (**25**) and some *n*-alkanes (**1**−**24**) and sterols (**85**−**96**) have a moderate concentration input. The source profile B has very low inputs of all of these components, but has higher concentrations of *n*-alkanes (**1**−**24**), UCM (**25**), pristane (**26**), phytane (**27**), perylene (**47**), PCBs (**66**−**74**), pesticides (**75**−**80**), and sterols (**85**−**96**), which are assumed to be of continental river origin (Ebro River). Finally the source profile C only has some high concentrations from some *n*-alkanes and very low concentrations from the other components. It corresponds to an unspecific background origin consistent with long-range transport of contaminants subjected to degradation and transformation processes during transport.

The three resolved source contribution profiles (matrix **R** in eq 1) according to the sampling site are shown in Figure 6. The highest contribution of source A was found in the

Gulf of Lion samples (Ty8, Ty14, Ty23, BC4, and BC6 in Figure 1). The highest contribution of source B was found in the Ebro area (BC7, C1, BC11, D1, D2, and D3 in Figure 1), and for source C was found in the open sea sampling sites (Ty17, Ty3, Ty27, BC12, BC14, BC15, TY19, BC8, and A2 in Figure 1).

All these results confirm previous findings with PCA and HCA methods, which both validates the former results and confirms the capacity of the proposed MCR-ALS approach by achieving multiple source apportionment (Figure 6). Interpretation of ALS resolved profiles is straightforward and provides an easy explanation of possible environmental sources of analyzed chemical compounds of samples collected in different geographical sites. In the present work, the MCR-ALS method is proposed for the first time as an exploratory tool for environmental data analysis.

Acknowledgments

Partial financial support was provided by the Spanish Plan for Research (Grant AMB 96-0926). The authors are indebted to Dr. J. Albaigés for his helpful discussions and encouragement during this work.

Literature Cited

(1) Bayona, J. M.; Farran, A.; Albaigés, J. *Mar. Chem*. **1989**, *27*, 79−104.
(2) Volkman, J. K. *Org. Geochem*. **1986**, *9*, 83−99.
(3) Sporstol, S.; Gjos, N.; Lichtenhaler, R. G.; Gustavsen, K. O.; Urdal, K.; Orels, F.; Skel, J. *Environ. Sci. Technol*. **1983**, *17*, 282−286.
(4) Zitko, V. *Mar. Pollut. Bull*. **1994**, *28*, 718−722.
(5) Naf, C.; Broman, D.; Pettersen, H.; Rolff, C.; Zebuhr, Y. *Environ. Sci. Technol*. **1992**, *26*, 1444−1457.
(6) Vogt, N. B.; Brakstadt, F.; Thrane, K.; Nordenson, S.; Krane, J.; Aamot, E.; Kolset, K.; Esbensen, K.; Steinnes, E. *Environ. Sci. Technol*. **1986**, *21*, 35−44.
(7) Wenning, R. J.; Paustenbach, D. J.; Harris, M. A.; Bedbury, H. *Arch. Environ. Contam. Toxicol*. **1993**, *24*, 271−289.
(8) Wenning, R.; Paustenbach, D.; Johnson, G.; Ehrlich, R.; Harris, M.; Bedbury, H. *Chemosphere* **1993**, *27*, 55−64.
(9) Tysklind, M.; Fängmark, I.; Marklund, S.; Lindskog; A.; Thanin, L.; Rappe, Ch. *Environ. Sci. Technol*. **1993**, *27*, 2190−2197.
(10) Evers, E. H. G.; Klamer, H. J. C.; Laane, R. W. P. M.; Govers, H. A. J.; *Environ. Toxicol. Chem*. **1993**, *12*, 1583−1598.
(11) Grimalt, J. O.; Olivé, J.; Gómez-Belinchon, J. I. *Intern. J. Environ. Anal. Chem*. **1990**, *38*, 305−320.
(12) Yunker, M. B.; Macdonald, R. W.; Veltkamp, D. J.; Cretney, W. J. *Mar. Chem*. **1995**, *49*, 1−50.
(13) Naes, K.; Oug, E. *Environ.Sci.Technol*. **1994**, *28*, 823−832.
(14) Henry, R. C.; Lewis, Ch. W.; Collins J. F. *Environ. Sci. Technol*. **1994**, *28*, 823−832.
(15) Readman, J. W.;Mantoura, R. F. C.; Llewellyn, C. A.; Preston, M. R.; Reves, A. V. *J. Environ. Anal. Chem*. **1986**, *27*, 29−54.
(16) Hopke, Ph. K. The Mixture Resolution Problem Applied to Airborne Particle Source Apportionment. Chemometrics in Environmental Chemistry. In *The Handbook of Environmental Chemistry, Volume 2, Part H;* Springer: Berlin, 1995; p 47.
(17) Tauler, R.; Smilde, A. K.; Kowalski, B. R. *J. Chemom*. **1995**, *9*, 31−58.
(18) Tauler, R. *Chemom. Intell. Lab. Syst*. **1995**, *30*, 133−146.
(19) Paatero, P.; Tapper, U. *Environmetrics*, **1994**, *5*, 111−126.
(20) Tolosa, I.; Bayona, J. M.; Albaiges, J. *Environ. Sci. Technol*. **1995**, *29*, 2519−2527.
(21) Tolosa, I.; Bayona, J. M.; Albaiges, J. *Environ. Sci. Technol*. **1996**, *8*, 2495−2503.
(22) Lipiatou, E.; Salliot, A. *Mar. Chem*. **1991**, *22*, 297−304
(23) Lipiatou, E.; Salliot, A. *Mar. Pollut. Bull*. **1991**, *22*, 297−304.
(24) Pirouette Multivariate Data analysis software for IBM PC systems. *Infometrix* Seattle, WA, 1992.
(25) MATLAB version 4.2. The MathWorks Inc.: 1994.
(26) Paatero, P.; Tapper, U. *Chemom. Intell. Lab. Syst*. **1993**, *18*, 183−194.
(27) Hopke, Ph. K. *Chemom. Intell. Lab. Syst*. **1991**, *10*, 21−43.
(28) Malinowski, E. *Factor Analysis in Chemistry*, 2nd ed.; Wiley: New York, 1991.
(29) Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617−632
(30) Borgen, O. S.; Kowalski, B. R. *Anal. Chim. Acta* **1985**, *174*, 1−26
(31) Windig, W.; Guilment, J. *Anal. Chem*. **1991**, *63*, 1425−1432.
(32) Windig, W. *Chemom. Intel. Lab. Syst*. **1992**, *16*, 1−16.
(33) Golub, G. H.; VanLoan, Ch. F. *Matrix Computations*, 2nd ed.; The John Hopkins University Press: Baltimore, MD, 1989.
(34) Wolfe, N. L.; Zepp, R. G.; Paris, D. F.; Baughman, G. L.; Hollis, R. C. *Environ. Sci. Technol*. **1977**, *11*, 1077−1081.
(35) Zoro, J. A.; Hunter, J. M.; Eglinton, G.; Ware, G. C. *Nature* **1974**, *247*, 235−247.
(36) Venkatesan, M. I. *Mar. Chem*. **1988**, *25*, 1−27.
(37) Ramdall, T. *Nature* **1983**, *306*, 580−582.
(38) Anderson, R. F.; Bopp, R. F.; Buesseler, K. O.; Biscaye, P. E. *Cont. Shelf Res*. **1988**, *8*, 925−946.
(39) Millot, C. *Oceanol. Acta* **1987**, *10*, 143−148.

Received for review March 13, 1997. Revised manuscript received August 26, 1997. Accepted August 30, 1997.[⊗]

ES970231E

[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1997.