4

Applications of PCA to the Monitoring of Hydrocarbon Content in Marine Sediments by Means of Gas Chromatographic Measurements

Mauro Mecozzi*, Marco Pietroletti, Federico Oteri and Rossella Di Mento
Laboratory of Chemometrics and Environmental Applications, ISPRA, Rome, Italy

1. Introduction

The application of Principal Component Analysis (PCA) in biochemical studies lies in the field of Chemometrics, the discipline which describes and applies statistical multivariate methods to the laboratory studies. PCA like Cluster Analysis (CA), belongs to the so called unsupervised pattern recognition methods, multivariate methods which can be applied to any data set without requiring or supposing any preliminary knowledge about the information present in the data (Massart & Kauffman, 1983; Brereton, 2003).

PCA has been also defined "a data reduction form" for its peculiar ability to reduce the dimension of an experimental data set without loosing the qualitative and quantitative information present (Brereton, 2003). In matrix notation, the PCA decomposition of a multivariate experimental data set including several samples and called X, is reported in the equation 1

$$X = SV' + E \tag{1}$$

where the S term is the score matrix, V' is the transposed loading matrix and E is the noise matrix . With respect to the original X (n-sample, t-variables) set, the dimension of the new matrices is changed; S has (n-sample, *p*) dimension, V has (p, t-variables) dimension and E only retains the same dimension of X obviously. The term " p " of S and V matrices represents the number of significant principal components or factors determined by PCA; they have the peculiar ability of describing a high fraction of the total variance (i.e. information) present in the X matrix and very important, the "p" dimension is always significantly lower than the " t " dimension of the original variables of the X matrix.

This data reduction ability of PCA is very helpful when large size of multivariate data sets have to be analyzed and interpreted. In common environmental monitoring studies PCA is applied in the analysis of discrete multivariate data when for instance, several sites with their pollutant loads have to be analysed and compared (Cicero et al., 2001; Conti & Mecozzi, 2008). However in environmental studies, the power of PCA becomes even more helpful when large size set of analytical signals such as GC chromatograms have to be

* Corresponding Author

analyzed. In fact, gas chromatography is a widespread technique for the monitoring of oil spills in terrestrial and marine environments (Wang et al., 1999) and in the case of marine sediments, gas chromatography tries to establish several aspects concerning total hydrocarbon content and distribution for testing homogeneity and or heterogeneity of pollutant loads and for identifying the sources of oil spills (Wang et al., 1999). In any case, this last task can be hardly obtained because any chromatogram is a multivariate sample where many hydrocarbons are usually present. A typical GC chromatogram, reported in Figure 1, is a data file with 2 columns, the acquisition time of the analytical signals and their detected intensities respectively. Here, the present hydrocarbons are identified by means of their retention time (i.e. the time corresponding to the maximum peak intensity).

The chromatogram of Figure 1 shows the presence of more than fifty hydrocarbons and in addition, the fast sampling signal causes the presence of a not negligible noise which corrupts the real intensity of signals (Mecozzi & Tomassetti, 2007; Kokaly et al., 2001). As a consequence, we can hardly perform a numerical and visual comparison of different chromatograms when we try to establish homogeneous or heterogeneous hydrocarbon compositions among samples as shown by the example of Figure 2.

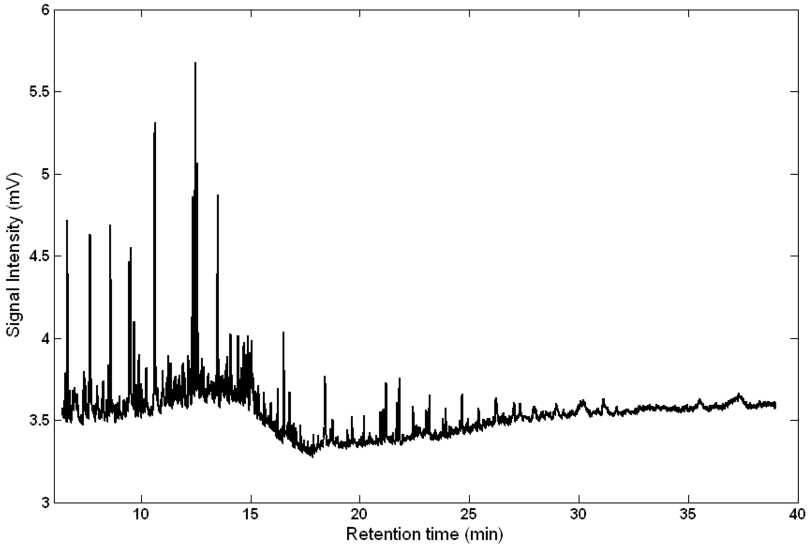

Fig. 1. Example of a GC chromatogram arising from the analysis of hydrocarbons extracted by a marine sediment. Any detected peak represents a hydrocarbon present in the sample.

According to Equation 1, PCA re-describes the starting X data set by means of a set of the new "p" variables (i.e. factors), which being significantly lower than the number of the original variables, allow to compare samples by means of simple two or three dimensional plots, using the score matrix. In addition, PCA examines the variables which determine similarity or dissimilarity among samples by means of the loading analysis. Loadings are the statistical weights of the original "t" variables of the X matrix and in the case of chromatographic data their analysis allows to identify the hydrocarbons which characterise any samples. This is a peculiar advantages of PCA with respect Cluster Analysis, that is known as a fast screening method to determine similarity in experimental data set but in any case, it allows neither to determine the statistical weight of the variable nor to study peculiar variables determining qualitative similarities and dissimilarities among samples (Brereton, 2003).

However, the application of PCA to large size data set requires some necessary preprocessing treatments so to avoid potential misinterpretation of its results. In fact, a GC data file. such as the chromatogram of Figure 1, consists of about 20.000 analytical signals and when we examine a data file including thirty or forty samples, the resulting X matrix has high data dimension and redundancy. This causes high time for PCA computation and analytical problems such as reduction of the signal to noise (S/N) ratio and baseline drift. The selection of proper preprocessing treatments of chromatograms can solve all these problems and supports the correct application of PCA to large size multivariate data.

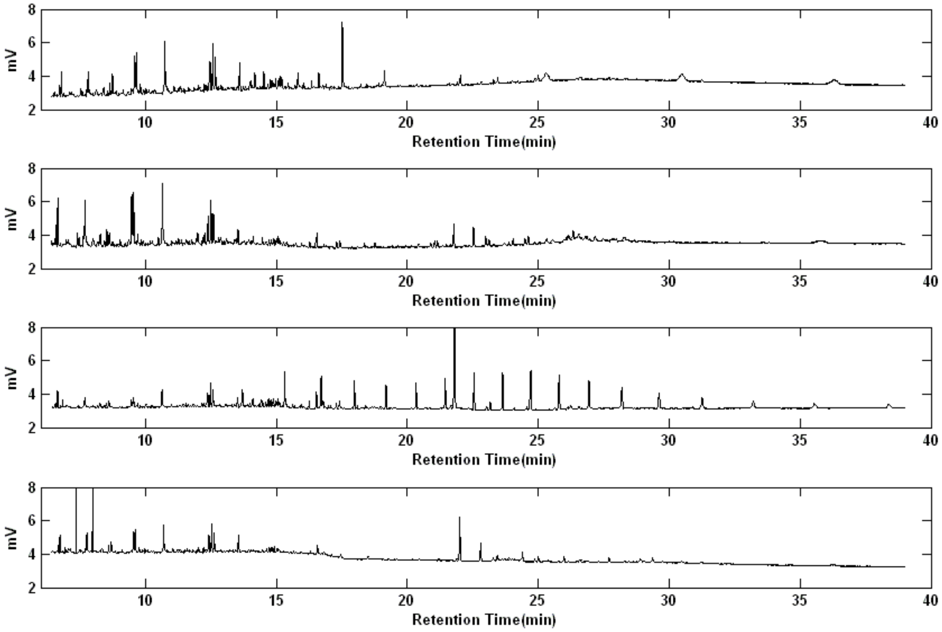

Fig. 2. GC chromatograms of hydrocarbons extracted by some sediments sampled along the Italian coasts. The simple visual examination of peak positions in the four plotted chromatograms shows how the related samples can be hardly compared for establishing similarities and dissimilarities of composition.

In this paper we discuss the application of PCA for performing the hydrocarbon monitoring in two different GC chromatographic sets. Our study takes into account all the steps for a correct application of PCA to high dimension chromatographic data files. The first set consists of 29 superficial sediments from two different areas along the coasts of Italian seas, seventeen from Venice lagoon (Adriatic sea) and twelve from Bagnoli (near Naples, Tyrrhenian sea) respectively; the second set consists of 39 subsamples of marine sediments coming from a sediment core taken in Antarctic sea.

The main purpose of PCA application is that to retrieve information hardly detectable by means of conventional methods of GC analysis of hydrocarbons in environmental studies.

2. Experimental section

This experimental study consists of five different steps; sampling of marine sediments, hydrocarbon extraction and purification from other lipid compounds present in marine sediments (Mecozzi et al., 2011), gas chromatographic analysis of the extracts, chemometric pretreatment of chromatograms and application of PCA. PCA was applied to the two different chromatographic data matrices including all the samples from the Italian coasts and the Antarctic sediment core.

2.1 Sampling of marine sediments

Marine sediment sampling from the Italian coasts was performed by a box corer, taking the upper 5 cm layer. Figure 3 reports the location of the two sampling areas along the Italian coasts. Samples were stored frozen at -25°C until chemical analysis.

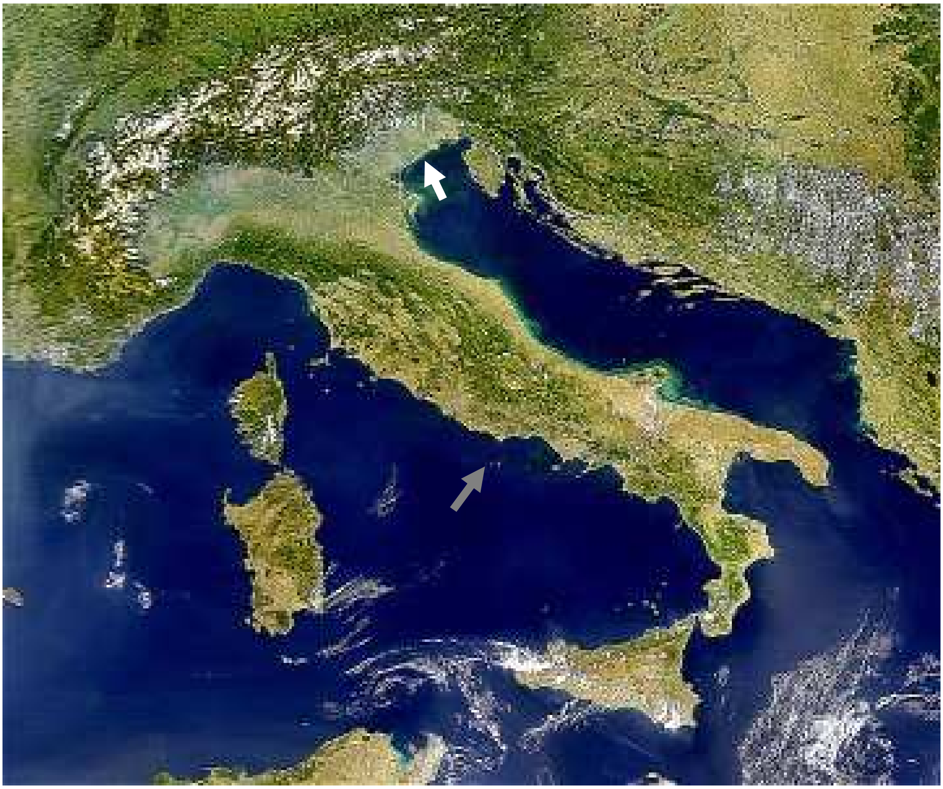

Fig. 3. Map of Italian coasts showing the two areas where surface sediments were sampled. The white arrows shows the area of the Venice Lagoon in Northern Adriatic sea and the grey one shows the area of Bagnoli near Naples in Tyrrhenian Sea.

The Antarctic sediment core was sampled in the B5/Y5 station (75° 04′ South, 164° 13′ East) in the Ross bay at 550 meter of depth. This area is characterised by an intense stratification of sediment and of biogenic organic materials. The sediment core was taken by means of dredge sampler and the core was stored frozen at -25°C until GC analysis.

2.2 Hydrocarbon extraction

Hydrocarbon content was extracted and purified by means of an ultrasound method developed in our laboratory (Mecozzi et al., 2011). Each sediment sample (20 g) was added with n-hexane (20 ml) and H_2O (40 ml) at pH 2 obtained by adding concentrated HCl. Sediment was sonicated in an ultrasound cleaning bath operating at 35 kHz for 20 minutes at room temperature. Then the supernatant was separated from sediment by centrifugation. The separation of the aqueous phase from the organic phase was performed in a separating funnel; then he organic phase was dried on anhydrous Na_2SO_4. This process was repeated other twice, the extracts joint together and the organic phase was concentrated under vacuum down to 1 ml of final volume for GC analysis.

2.3 Gas chromatographic analysis

The determinations of hydrocarbons extracted by marine sediments were performed using a Carlo Erba (Milano Italy) instrument with flame ionization detector. The apparatus was equipped with a capillary GC Column Therm 1 (Thermo Scientific Milano Italy), 30 m length, i. d. 0.22 mm. Experimental conditions were injector 320°C, FID detector 360°C and the introduction was performed in spleatless mode (one minute). The temperature program used for chromatographic separation of hydrocarbons was 70°C for four minutes, thermal gradient 15°C min^{-1} to 340°C; This temperature was finally held for fourteen minutes. Chromatograms were saved as ASCII files for any further elaboration.

2.4 Chemometric pretreatments of chromatograms prior to PCA application

2.4.1 Improvements of analytical quality data and reduction of computation time

Handling of large data set prior to PCA application requires the preliminary solution of several drawbacks; in fact, the high frequency sampling of analytical signals produces data redundancy, high time of computation, with in addition analytical drawbacks such as reduction of the signal to noise (S/N) ratio and baseline drift (Christensen and Tomasi (2007). The same authors suggested several chemometric procedures for reducing these effects prior to apply PCA to GC data; with this aim, an in house MATLAB (Natik, USA) routine was applied to any collected chromatogram. In the appendix we report a MATLAB routine according to the algorithms described by Christensen and Tomasi (2007). Figure 4 reports an example of this approach. After this pretreatment, GC chromatograms were saved again as ASCII files.

2.4.2 Standardisation of the GC data set

Standardisation, also called scaling, is another fundamental step prior to PCA application, necessary for reducing the effect of the different magnitude of intensity variations in the case of multivariate data, causing uncorrected determination of the total variance of the data system (Brereton, 2003; Wang et al., 1999; Noda, 2008).

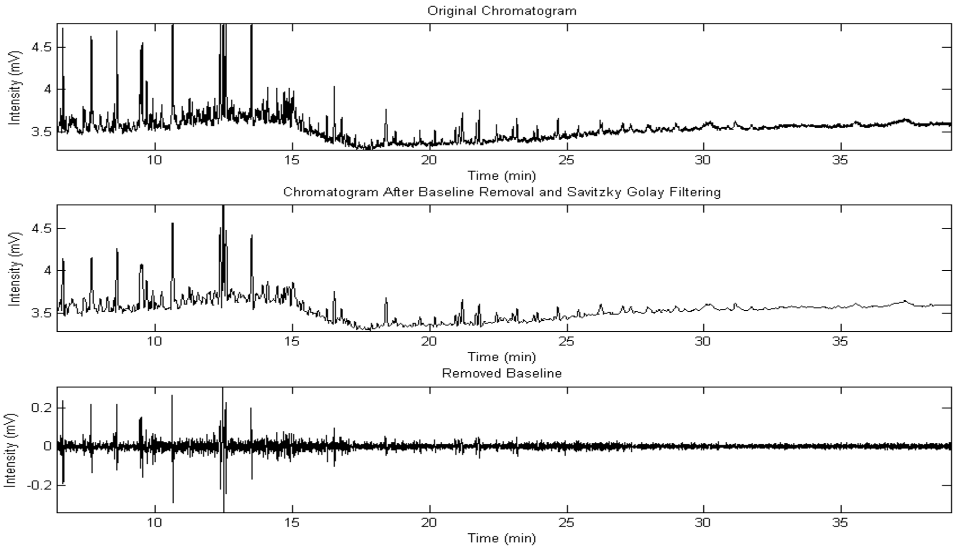

Fig. 4. Example of chemometric pretreatment of a GC chromatogram., Original chromatogram, upper plot; chromatogram with data redundancy reduction, smoothed signals and background correction, middle plot; bottom plot, removed baseline.

In environmental monitoring, where PCA is often applied to study the distribution of pollutant loads, a common scaling technique is autoscaling; given the Y column vector to be included in the X matrix and having n-sampled analytical signals, autoscaling performs the data transformation according to

$$Y_{ias} = (Y_i - Y_M)/\sigma \qquad (2)$$

where Y_i , Y_{ias} , Y_M and σ are the original value i_{th} value, its autoscaled term, the average value of the Y vector and standard deviation of the Y vector respectively. After autoscaling, any new Y series to be included in the X matrix has mean value 0 and variance value 1.

This is a very powerful approach to reduce the effect of different size ranges on the total variance of discrete data set but when applied to other types of variables such as the cases of analytical signals, autoscaling has a marked drawback. In fact, digitised files of spectroscopic and chromatographic data generally consist of several thousands of signals sampled with high frequency acquisition. In this case autoscaling can often produce the enhancement of noise depending on its division by a small value of standard deviation (Noda, 2008; Kokalj et al., 2011).

Other scaling techniques are available for solving the disadvantage originating from autoscaling. In the mean centred technique data are scaled according to

$$Y_{imc} = (Y_i - Y_M) \qquad (3)$$

where Y_{imc} is mean centred scaled value of the Y series while Y_i and Y_M are the same meaning of the equation 2.

Normalization scaling consists of transforming data according to

$$Y_{inorm} = (Y_i - Y_{min})/(Y_{max} - Y_{min}) \tag{4}$$

where Y_{inorm}, Y_{min} and Y_{max} are the normalized Y_i term, the minimum and the maximum values of the Y series respectively. After normalization, all the Y vectors range between 0 and 1.

Pareto scaling is a technique proposed by the Italian economist Vilfredo Pareto (Noda, 2008); it consists of the division of the Y series values by the square root of its standard deviation according to

$$Yi_p = Y_i/\sqrt{\sigma} \tag{5}$$

where Y_{ip} is the Pareto scaled of the original Yi value and σ has the same meaning of equation 2.

Any scaling technique produces different effects on the quality of analytical signals so that the selection of the opportune scaling needs a carefully evaluation of the produced results. We report examples of application of all the above scaling methods in Figures 5 and 6 so to support the selection of the most appropriate methods prior to PCA application to GC data. With respect to the original chromatogram, autoscaling causes baseline drift with negative analytical signals and in addition, noise is enhanced in some zones of the chromatogram as shown by the example of Figure 5 (middle plot).

Mean centred scaling causes a baseline drift with negative analytical signals as well, though it does not cause a S/N ratio reduction as observed for autoscaling instead (Figure 5, bottom plot).

Normalization and Pareto scaling techniques do not cause negative baseline drifts and evident noise enhancements (Figure 6) so that we recommend to apply one of these as scaling pretreatments. These techniques can be applied by means of a common spreadsheet such as Excel for Windows. In any case, in the Appendix section we report two ad hoc routines written in MATLAB language for applying the above scaling techniques.

2.5 Application of PCA to gas chromatographic data set

PCA was applied to GC chromatograms by an in house routine written in MATLAB (Natik, Wi, USA, ver 5.0) language according to the singular value decomposition algorithm described by Geladi (2002). The list of the routine is reported in the Appendix section.

2.6 Chemical reagents

All the chemical reagents used for the experimental work were of analytical reagent grade (Carlo Erba, Milan, Italy) and only ultrapure MilliQ water was used for any chemical treatments of samples.

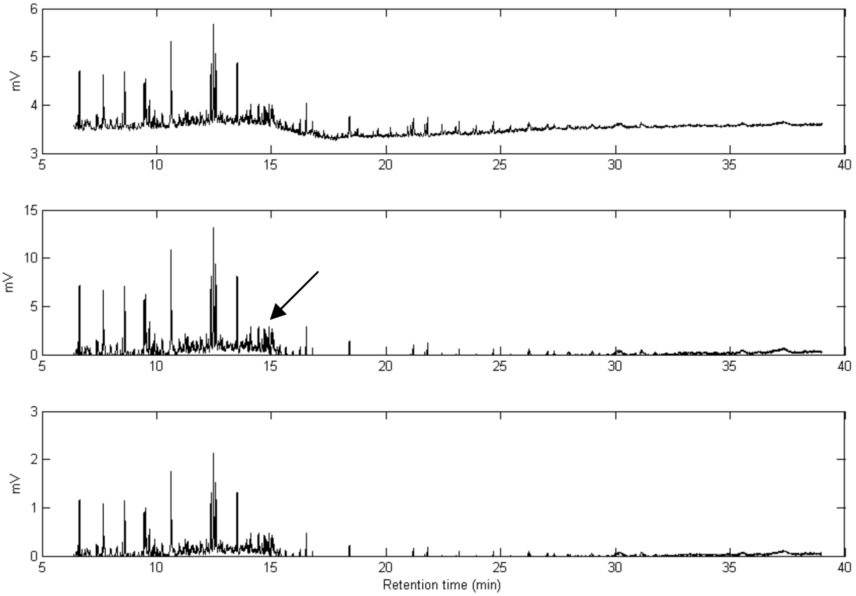

Fig. 5. Scaling methods applied to GC set. Conventional chromatogram, upper plot ; autoscaling, middle plot; mean centred scaling, bottom plot. The arrow shows a case where autoscaling increases noise with respect to the original plot.

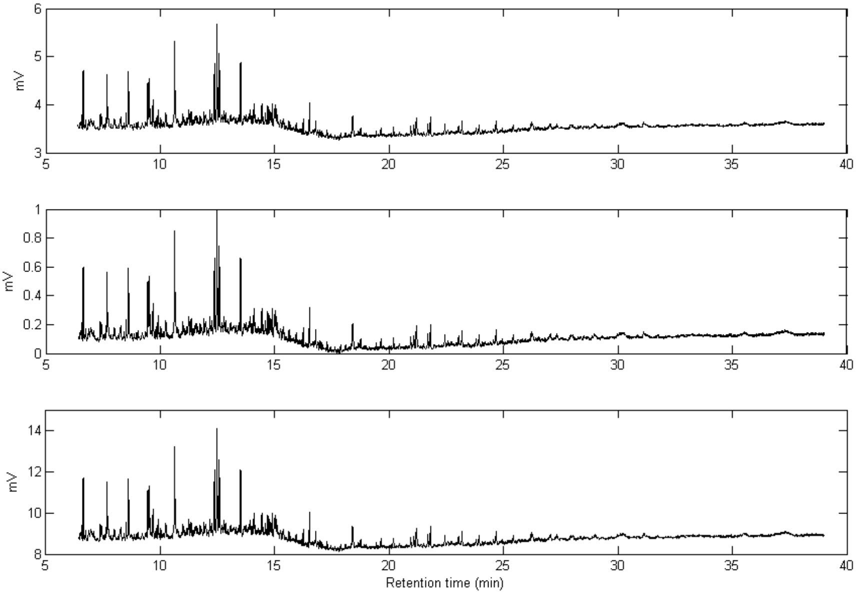

Fig. 6. Plots of original chromatogram (upper plot), normalization scaling chromatogram (middle plot) and Pareto scaling chromatogram (bottom plot).

3. Results and discussion

3.1 Application of PCA to hydrocarbon analysis in sediments from two areas of Italian coasts

Figure 7 reports the score plot of the first vs. the second factor obtained by PCA application to the GC chromatograms of superficial sediment samples taken along the coasts of Adriatic and Tyrrhenian sea. These two factors extracted by PCA explain the 90.7 % of a total variance of the chromatographic data set. This very high fraction of information, retained in two factors only, is an impressive example of PCA ability as "data reduction form"; now, the visual comparison of GC samples is possible by means of a simple two-dimensional plot depending on the reduction of the starting 20.000 variables (i.e. the retention times of hydrocarbons) to the two PCA factors.

The clustering of samples determining homogeneity and heterogeneity among samples is also evident and does not require further multivariate methods such ad discriminant analysis to investigate the classification of samples. Though these samples come from different seas and areas, some samples of the two areas have comparable hydrocarbon compositions as results from several VL and BG samples present in a same cluster, while samples of the Bagnoli area show different hydrocarbon compositions. This result means that the contributions of several biogenic (i.e. natural) and anthropogenic hydrocarbons can make sometimes comparable even sediments from different areas such as the two seas. These results can be hardly retrieved by the visual examination of the 29 chromatographic plots.

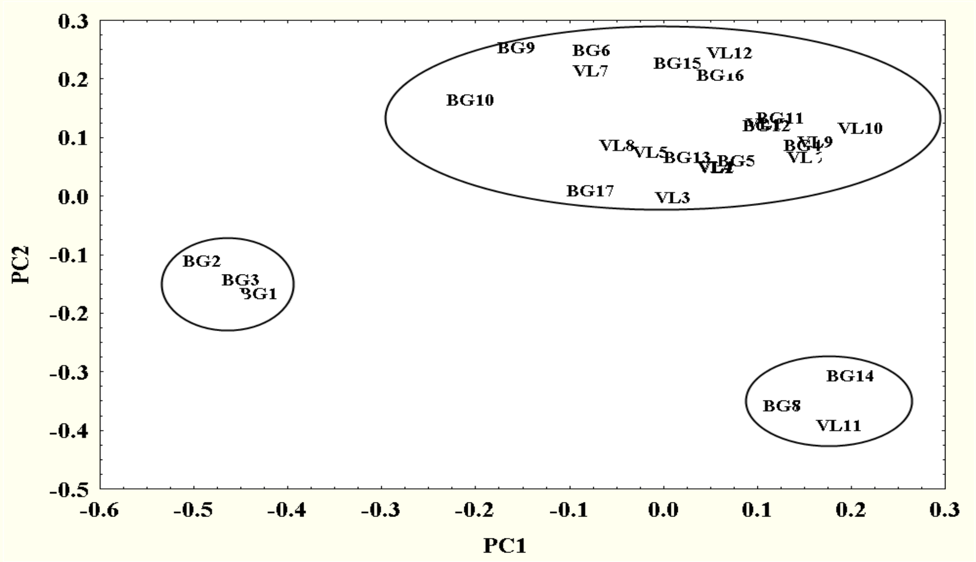

Fig. 7. Score plot of the first (PC1) vs. the second (PC2) factor from PCA applied to GC chromatograms of the Venice lagoon in Northern Adriatic (VL) and Bagnoli (BG) in Tyrrhenian sea. The two factors explain the 83.9 % and the 6.8 % respectively of the total variance. The ellipses are arbitrary and show the three different clusters.

However PCA can give additional information concerning the qualitative composition of samples because loading analysis can detect the hydrocarbon characteristics determining the similarities and dissimilarities observed in Figure 7.

The loading plot of the first factor (Figure 8) shows the generally high variability present in the hydrocarbon distribution of environmental samples as this factor explains the 83.9% of the total variance. Moreover, Figure 8 shows allows to retrieve characteristics concerning the hydrocarbon distribution of these samples. Pristane and phytane are two peculiar hydrocarbons able to characterise the biogenic and the anthropogenic sources present in environmental samples. In fact, pristane is a hydrocarbon typical of biogenic sources whereas phytane is a hydrocarbon typical of anthropogenic sources (Wang et al., 1999; Mecozzi et al., 2008; Duan et al., 2010). In this loading plot, pristane is negligible (retention time 15.5 minutes) whereas phytane is present (retention time 16.2 minutes in Figure 8, upper plot). In addition, the wax hydrocarbons (i.e. number of carbon higher than 24) which are also typical of biogenic sources (Wang et al., 1999; Duane et al., 2010; Ibbotson and Ibhadon, 2010; Ahad et al., 2011), are absent as shown by the negligible presence of chromatographic peaks with retention time higher than 20 minutes (Mecozzi et al., 2011). So the first loading plot describes the anthropogenic feature of the examined samples.

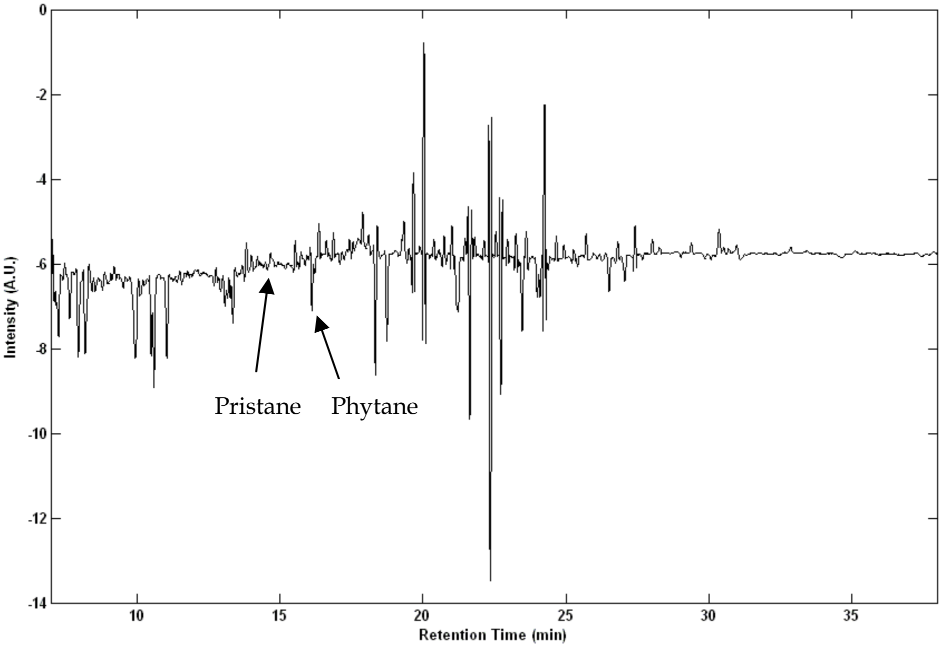

Fig. 8. Loading plot of first factor for data from the two areas from the Venice lagoon and Bagnoli near Naples. The arrows show the position of pristane (negligible) and phytane (present) in the corresponding chromatogram plot.

The loading plot of the second factor (Figure 9) , though explaining about the 7% of the total variance only, shows that samples in the upper cluster of Figure 7 are characterised by little concentration changes of some specific hydrocarbons related to biogenic hydrocarbon sources. In fact, with respect to the loading plot of Figure 8, here several linear hydrocarbons with carbon number higher than 24 are present and this is a marker of biogenic hydrocarbons (Wang et al, 1999; Duane et al., 2010). Obviously, due to the heterogeneity of the hydrocarbon composition, PCA can not specify the concentration changes of a single hydrocarbon, but in any case, it is relevant that we can compare samples of different origins solving the problem related to the general lack of methods to compare regional differences in areas submitted to potential hydrocarbon spills (Fraser at al., 2008).

Another interesting and useful advantage of using PCA in GC monitoring data consists of its support to the application of another well diffused unsupervised pattern recognition method such as Cluster Analysis. According to its name, CA performs the classification of data by identifying clusters of data having relevant similarities and for this purposes, it uses the multivariate distance among samples (Massart and Kaufmann, 1983).

CA is considered a fast screening method to perform exploratory data analysis though it does not identify the variables which determine similarity and or dissimilarity among samples; this remains a peculiar ability of PCA (Figures 8 and 9).

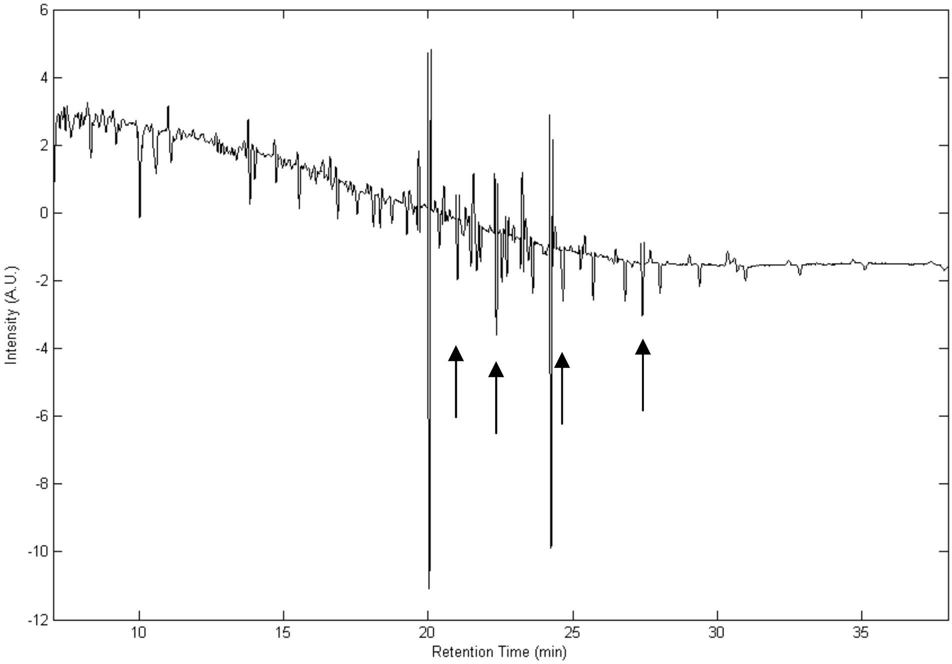

Fig. 9. Loading plot of second factor for data from the two areas from the Venice lagoon and Bagnoli near Naples. The arrows show the presence of some linear high molecular weight hydrocarbons, with more than 24 carbon atoms, typical of biogenic sources.

However, the application of CA to samples with over than 20.000 variables such as the case of the GC data is almost impossible due to computational and collinearity problems among variables (Massart and Kaufman, 1983). Conversely when CA is applied by means of the PCA scores, we have many peculiar advantages because this approach requires a small number of uncorrelated factors only while it does not require the use of specific distance such as the Mahalanobis one (Massart and Kaufman, 1983). This approach reported in Figure 10, shows that samples are clustered in a perfect agreement with Figure 7 obviously and now, by means of the data reduction of PCA, we can apply CA for estimating the percent of similarity existing among samples.

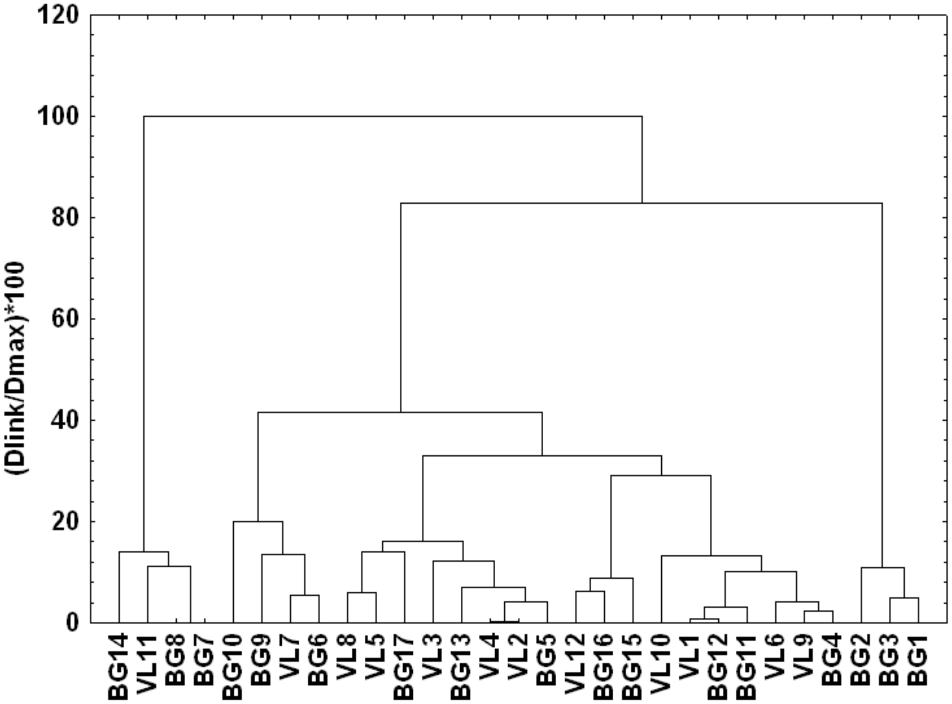

Fig. 10. Cluster Analysis of GC data performed by means of PCA scores. The ratio (Dlink/Dmax)* 100 of the ordinate axis is the quantitative measurements of the dissimilarity among samples.

3.2 Application of PCA to hydrocarbon analysis of sediment samples from an Antarctic core

The application of PCA to the chromatographic data set of an Antarctic sediment core (Figure 11) gives even more peculiar results with respect to those obtained in the previous section. Being Antarctic continent uncontaminated, we can suppose reasonably that hydrocarbons present in sediment core samples depend on biogenic contribution essentially with negligible anthropogenic contributions. If so, the hydrocarbon composition changes observed along the sections of the Antarctic sediment core should have a qualitative homogeneous composition depending on the biogenic contributions. As a consequence, the observed quantitative changes should depend on the natural stratification events only. The results reported in the score plot of Figure 11 supports this hypothesis.

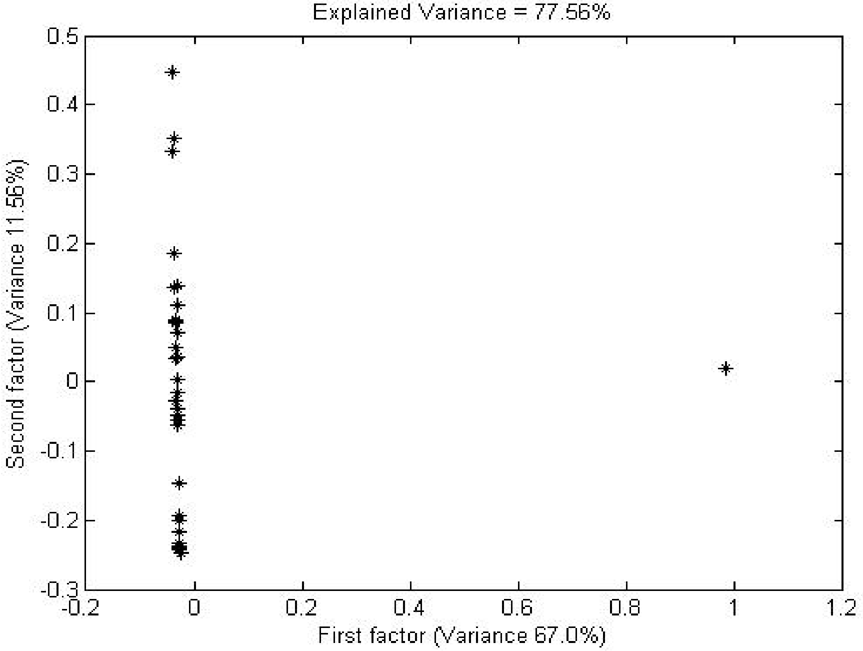

Fig. 11. Score plot of the first vs. the second factor from PCA applied to GC chromatograms of the Antarctic sediment core. The two factors explain the 67.0 % and the 11.56 % respectively of the total variance.

The first factors explains the 67.0 % of the total variance and the score values have an almost constant value while the positions of the samples changes with the scores of the second factor explaining the 11.56 % of the variance. The loading analysis reported in Figure 11, gives many details for the clarification of these findings. In the first factor, it is evident the presence of a significant hydrocarbon peak at high molecular weight (retention time close to 35 minutes) assigned to the linear hydriacrbon with 38 carbon atom number. This is a wax hydrocarbon, typical of biogenic contributions arising from the degradation of living cells (Duane et al., 2010).

PCA confirms the supposed prevalence of biogenic contributions for these samples depending on prevalent presence of the biogenic linear hydrocarbon with 38 carbon number, suggesting a significant homogeneous composition mostly governed by the natural stratification of sediments as well. In addition, if the hydrocarbon distribution along the sections core is determined by the natural stratification of sediments only, we can suppose that it is governed by time. In this case, we could test the hypothesis of the time depending relationship between stratification of hydrocarbon distribution in sediments by means of the autocorrelation function, a typical approach for time series analysis (Brereton, 2003). In fact, autocorrelation is a tool for studying time trend and periodicity present in an univariate data set according to the regressive model

$$Y_{t+1} = mY_t + \ \text{cost} \qquad\qquad t = 1, 2, \ldots\ldots n \qquad (6)$$

Autocorrelation has an easy application to univariate time series data but its application to multivariate data such chromatographic ones can be performed after a PCA data reduction, under the condition that its first factor explains a high percent of the total variance (Brereton, 2003). In the case of the Antarctic core samples this condition is fulfilled (i.e. 67% in the first factor) and the first factors can be considered as an univariate time series. So we can examine our data by the autocorrelation method using the score values of the first factor.

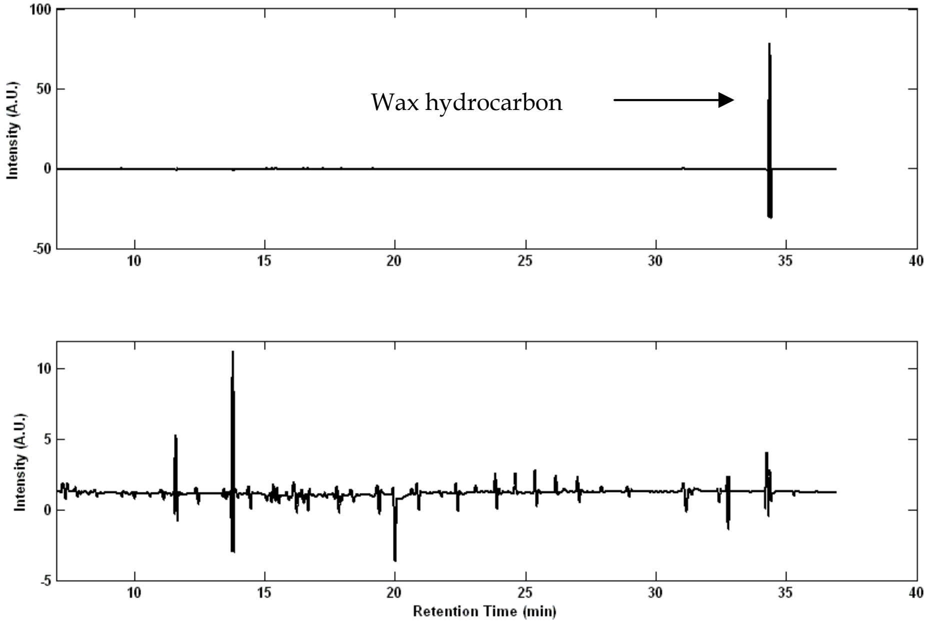

Fig. 12. Loading plot of the first (upper) and second (bottom) factors related to data of the Antarctic sediment core.

We report the result of this time series PCA application in Figure 13. The time descending trend of hydrocarbon distribution, depending on the natural stratification of sediments only, is clearly supported by the shape of the autocorrelation plot. On the base of this finding, we can verify that the hydrocarbon distribution shows a time trend depending on its biogenic contributions, because if anthropogenic sources were also present, we should observe a more irregular vertical profile and not the time trend supposed by Figure 11 and clearly confirmed by Figure 13.

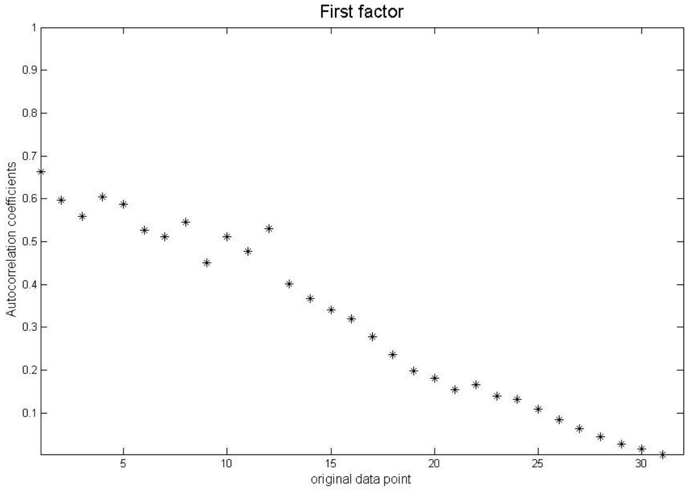

Fig. 13. Autocorrelation plot of the first factor of PCA applied to the hydrocarbon data distribution. The abscissa axis corresponds to the number of the sediment sections of the Antarctic core, while the number on the ordinate axis corresponds to the value of the autocorrelation function determined by the first PCA factor.

4. Conclusion

In this study, we have presented all the aspects of pretreatment, scaling and use related to the application of PCA to complex multivariate chromatographic data set coming from environmental studies. As far as data pretreatment concerns, we stress the importance of data redundancy reduction, signal to noise improvement and data scaling. For this latter aspect, we have evidenced the peculiar advantages given by normalization and Pareto scaling techniques with respect to the most applied autoscaling technique. When applied to two specific cases of environmental studies PCA allows to retrieve much more information than that obtained by the conventional visual examination of GC chromatograms. Both case of studies show the power of PCA for explorative data analysis in chromatography and in addition, its ability as “data reduction form ”supports the use of other statistic and complimentary techniques such CA and Time Series Analysis in the interpretation and verification of the environmental results.

5. Acknowledgments

This study has been jointly supported by the research project “ASTRA” financed by the Ente Nazionale Idrocarburi (ENI), Milan, Italy and by the National Research Antarctic Program (PROGDEF09_125) financed by the Italian Minster of the University and Research.

6. Appendix

6.1 MATLAB routine for performing reduction of data redundancy, smoothing and baseline correction and removal

```
function [d,g]=gcpretreatment(chromatogram,factored);
```

```
% Routine for data redundancy reduction, Savitzky Goaly filtering and  baseline removal in
chromatograms
% Ref. "Practical Aspects of Chemometrics for Oil Spill Fingerprinting"
% J.H. Chrinstensen and G. Tomasi, J. Chromatography A, 1169 (2007) 1-122
% chromatogram  is  a data file with X (time),Y(mV) in ASCII format
% factored is a number specifying the entity of the data reduction
% for N  number of X,Y couples of signals, factored=2 gives  a final an N/2 length data file
% "g" is the reduced chromatogram
% "d" is the removed baseline
% “g” has to be saved  as ASCII file for further elaboration by PCA
% Data redundancy reduction
a=chromatogram(:,1);
b=chromatogram(:,2);
a(1:factored:length(a))=[];
b(1:factored:length(b))=[];
% Formation of reduced chromatogram (matrix)
c=[a,b];
%
% Baseline removal
rid=c(:,2);
% determination of baseline "d"
for i=1:length(rid)-1
   if i==1
   d(1)=rid(1);
   end
d(i)=rid(i+1)-rid(i);
end
d=[d(1),d];
f=rid-d';
%
% Savitzki Golay  filtering (third order function 17 points)
g=sgolayfilt(f,3,35);
subplot(3,1,1)
plot(chromatogram(:,1), chromatogram(:,2),'k')
title('Original Chromatogram')
xlabel('Time (min)')
ylabel('Intensity (mV)')
axis([min(chromatogram(:,1))  max(chromatogram(:,1))  min(g)  max(g)])
subplot(3,1,2)
plot(a,g,'k')
title('Chromatogram After Baseline Removal and Savitzky Golay Filtering')
xlabel('Time (min)')
ylabel('Intensity (mV)')
axis([min(a) max(a) min(g) max(g)])
subplot(3,1,3)
plot(a,d,'k')
title(' Removed Baseline ')
xlabel('Time (min)')
ylabel('Intensity (mV)')
```

```
axis([min(a) max(a) min(d) max(d)])
```

6.2 MATLAB routine for performing normalisation scaling

```
function [b]=norma(dataset);
% Routine for normalization of a spectral or chromatographic data samples
% Data matrix is column wise; each column corresponds to one sample
% dataset is the ASCII files of data to be normalized
% Files are uploaded as ASCII file
[m,n]=size(dataset);
for j=1:m
for i=1:n
b(j,i)=(dataset(j,i)-min(dataset(:,i)))/(max(dataset(:,i))-min(dataset(:,i)));
end
end
```

6.3 MATLAB routine for performing Pareto scaling

```
function [b]=paretoscaling(dataset);
% Routine for scaling of a spectral or chromatographic data samples
% by the Pareto approach
% Data matrix is column wise; each column corresponds to one sample
% dataset is the ASCII files of data to be normalized
[m,n]=size(dataset);
for j=1:m
for i=1:n
b(j,i)=(dataset(j,i))/sqrt(std(dataset(:,i)));
end
end
```

6.4 MATLAB routine for performing PCA

```
function [scores,loadings,varpercent]=pcawp(x);
% Principal Component Analysis (PCA)according to the algorithm Singular Value
% described by P. Geladi in Calculating Principal Component Loadings and Scores
% ISBN 91-7191-083-2, Umea, Sweden
% the “x” file is the data file after all the pretreatments
% Matrix r*c whit "r" rows (samples) and "c" columns (variables) can be analysed
% use the "save filename.txt a –ascii -tabs" instruction for saving files
% varpecent is the file with percent of variance explained by all the factors
[u,d,v]=svd(x);
% Determination of explained variance retained by each factor
l=d.*d;
varpercent=diag(l./trace(l))*100
scores=u*d;
loadings=v;
```

7. References

Ahad, J.M.E.; Ganeshram, R. S.; Bryant, C. L.; Cisneros-Dozal, L.; Ascough, P. L.; Fallick, A.E. & Slater, G.F. (2011). Sources of n-alkanes in un urbanized estuary: Insight from molecular distributions and compound-specific stable and radiocarbon isotopes. Marine Chemistry, Vol. 126, No. 1, 239-249, ISSN 0304-4203

Brereton, R. (2003). Chemometrics, Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons, West Sussex Po, UK, ISBN 0-471-48977-8

Christensen, J. H. & Tomasi, G. (2007). Practical Aspects of Chemometrics for Oil Spill Fingerprint. Journal of Chromatography A, Vol. 1169, No. 1-2 , 1-22, ISSN 0021-9673

Cicero, A.; Mecozzi, M.; Morlino, R. ; Pellegrini, D. & Veschetti E. (2001). Distribution of Chlorinated Organic Pollutants in Harbor Sediments of Livorno (Italy): A Multivariate Approach to Evaluate Dredging Sediments. Environmental Monitoring & Assessment, Vol. 71, No.3, 297-316, ISSN: 0167-6369

Conti, M. E. & Mecozzi, M. Multivariate Approaches in Biomonitoring Studies in 'Biological Monitoring: Theory and Application' , M.E. Conti Editor, (2008). Southampton, UK, WIT PRESS, ISBN: 978-1-84564-002-6

Duane, M. ; He, K. & Liu, X, (2010). Characteristics and Source Identification of Fine Particulate n-alkanes in Beijing, China. Journal of Environmental Sciences, Vol. 22, No. 7, 998-1005, ISSN 1001-0742

Fraser. G.S.; Ellis, J. & Hussain, L. (2008). An International comparison of Governmental Disclosure of Hydrocarbon Spills from Offshore Oil and Gas Installation. Marine Pollution Bulletin, , Vol. 56, No.1, 9-13, ISS 0025-326X

Geladi P. (2002). Calculating Principal Component Loadings and Scores in MATLAB, ISBN 91-7191-083-2, Umeå, Sweden

Kokaly, M.; Rihtarič, M. & Kreft, S. (2011). Commonly Applied Smoothing of IR Spectra Showed Unappropriate for the Identification of Plant Leaf Samples. Chemometrics and Intelligent Laboratory Systems, Vol. 108, No.2, 154-161, ISS 0169-7439

Ibbotson, J. & Ibhadon. A.O. (2010). Origin and analysis of aliphatic and cyclic hydrocarbons in northeast United Kingdom coastal marine sediment. Organic Geochemistry, Vol. 60, No. 10, 1136-1141, ISSN0025-326X

Massart, D. L. & Kaufmann, L. 1989. The Interpretation of Analytical Chemical Data by the use of Cluster Analysis. Robert E. Krieger Publishing Company, Malabar, Florida, USA, ISBN 0-89464-358-4

Mecozzi, M. & Tomassetti, P. (2007). Handling of a Large Data Set: Application of Time Series Analysis to Oceanographic Studies. International Journal of Environment & Health, Vol. 1, No.6, 347-359, ISSN 1743-4955

Mecozzi M.; Scarpiniti, M.; Ragosta, E.; Pietroletti, M. & Di Mento, R. (2008). Proposal for a deconvolution procedure for the gas chromatographic estimation of pristane and phytane in marine sediments. International Journal of Environment & Health, Vol.3, No.1, 126-138, ISSN 1743-4955

Mecozzi, M.; Pietroletti, M.; Mattiello, S.; Moscato, F. & Oteri, F. (2011). Proceedings of 17° International Symposium on Separation Sciences, Cluj, Romanian, 5-9 September. ISBN 978-973-133-981-8

Noda, I. (2008). Scaling Techniques to Enhance two-Dimensional Correlation Spectra. Journal of Molecular Structure, Vol. 883-884, No.1, 216-227, ISS 0022-2860/S

Wang, Z.; Fingas, M. & Page, D.S. (1999). Oil Spill Identification. Journal of Chromatography A, Vol. 843, No. 1-2, 369-411, ISS 0021-9673(99)00120-X

ISBN 978-953-51-0182-6
Hard cover, 230 pages
Publisher InTech
Published online 07, March, 2012
Published in print edition March, 2012

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821