Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
SCIENZE CHIMICHE

CICLO XXII

COORDINATORE Prof. GILLI Gastone

STUDY OF MULTICOMPONENT CHROMATOGRAMS USING A CHEMOMETRIC APPROACH: CHARACTERIZATION OF THE ORGANIC FRACTION OF ATMOSPHERIC AEROSOL

Settore Scientifico Disciplinare CHIM/01

Dottorando
Dott. MERCURIALI Mattia

Tutore
Prof.ssa PIETROGRANDE Maria Chiara

Anni 2007/2009

Alla mia famiglia

Contents

List of Papers . . . 1

1 Introduction 2

1.1 Complex Mixtures: Multicomponent Chromatograms . . . 3

1.2 Signal Processing . . . 4

2 AutoCoVariance Function 6

2.1 Theory . . . 6

2.1.1 Poissonian-Type Chromatogram . . . 9

2.1.2 Ordered-Type Chromatogram . . . 10

2.2 Homologous Series . . . 12

2.3 Linearization . . . 14

2.4 Odd/Even Prevalence . . . 18

3 Particulate Matter 23

3.1 Organic Fraction of PM . . . 25

3.2 Markers and Source Apportionment . . . 27

3.2.1 n-Alkanes . . . 28

3.2.2 Carboxylic Acids . . . 30

4 Results and Discussion 33

4.1 N-alkanes . . . 34

4.1.1 Application to samples of known origin 34

4.1.2 Application to real samples of PM 39

4.2 Carboxylic acids 43

4.2.1 Low Molecular Weight (LMW) dicarboxylic acids 43

4.2.2 N-alkanoic acids 48

5 Conclusions **50**

6 Future Perspectives **52**

Appendix **55**

Appendix A 55

Appendix B 55

List of Figures **61**

List of Tables **61**

Bibliography **68**

Acknowledgements **78**

UNIFE FORMAZIONE POSTLAUREA

Registrazione modulo Dichiarazione di conformità

MODULO INVIATO CORRETTAMENTE

Consegnare la copia stampata e debitamente firmata all'Ufficio Dottorato e Alta Formazione in via Scienze 41b Ferrara

Io sottoscritto Dott. (Cognome e Nome)
Mercuriali Mattia
nato a
Cotignola
Provincia
RA
il giorno
11/01/1982
Your E-Mail Address
mattia.mercuriali@gmail.com
avendo frequentato il corso di Dottorato di Ricerca in:
Scienze Chimiche
Ciclo di Dottorato
XXII
Titolo della tesi in Italiano
STUDIO DI CROMATOGRAMMI MULTICOMPONENTE USANDO UN APPROCCIO CHEMIOMETRICO: CARATTERIZZAZIONE DELLA FRAZIONE ORGANICA DELL'AEROSOL ATMOSFERICO
Titolo della tesi in Inglese
STUDY OF MULTICOMPONENT CHROMATOGRAMS USING A CHEMOMETRIC APPROACH: CHARACTERIZATION OF THE ORGANIC FRACTION OF ATMOSPHERIC AEROSOL
Titolo della tesi in altra Lingua Straniera
Tutore - Prof:
Prof.ssa Maria Chiara PIETROGRANDE
Settore Scientifico Disciplinare (SSD)
CHIM/01
Parole chiave (max 10)
multicomponent chromatogram, chemometric approach, EACVF, PM
Consapevole - Dichiara
CONSAPEVOLE --- 1) del fatto che in caso di dichiarazioni mendaci, oltre alle sanzioni previste dal codice penale e dalle Leggi speciali per l'ipotesi di falsità in atti ed uso di atti falsi, decade fin dall'inizio e senza necessità di alcuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni; -- 2) dell'obbligo per l'Università di provvedere al deposito di legge delle tesi di dottorato al fine di assicurarne la conservazione e la consultabilità da parte di terzi; -- 3) della procedura adottata dall'Università di Ferrara ove si richiede che la tesi sia consegnata

dal dottorando in 4 copie di cui una in formato cartaceo e tre in formato .pdf, non modificabile su idonei supporti (CD-ROM, DVD) secondo le istruzioni pubblicate sul sito : http://www.unife.it/dottorati/dottorati.htm alla voce ESAME FINALE – disposizioni e modulistica; -- 4) del fatto che l'Università sulla base dei dati forniti, archivierà e renderà consultabile in rete il testo completo della tesi di dottorato di cui alla presente dichiarazione attraverso l'Archivio istituzionale ad accesso aperto "EPRINTS.unife.it" oltre che attraverso i Cataloghi delle Biblioteche Nazionali Centrali di Roma e Firenze. --- DICHIARO SOTTO LA MIA RESPONSABILITA' --- 1) che la copia della tesi depositata presso l'Università di Ferrara in formato cartaceo, è del tutto identica a quelle presentate in formato elettronico (CD-ROM, DVD), a quelle da inviare ai Commissari di esame finale e alla copia che produrrò in seduta d'esame finale. Di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi; -- 2) di prendere atto che la tesi in formato cartaceo è l'unica alla quale farà riferimento l'Università per rilasciare, a mia richiesta, la dichiarazione di conformità di eventuali copie; -- 3) che il contenuto e l'organizzazione della tesi è opera originale da me realizzata e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura civile, amministrativa o penale e sarà da me tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi; -- 4) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati, non è oggetto di eventuali registrazioni di tipo brevettale o di tutela. --- PER ACCETAZIONE DI QUANTO SOPRA RIPORTATO

Firma Dottorando

Ferrara, lì _____02/03/2010_____________________

Firma del Dottorando ________________________________

Firma Tutore

Visto: Il Tutore

Si approva

Firma del Tutore _______________________________________

FORMAZIONE POSTLAUREA

Ufficio Dottorato di Ricerca - Ufficio Alta Formazione ed Esami di Stato - IUSS

List of Papers

- Paper I - Signal Processing of GC-MS Data of Complex Environmental Samples: Characterization of Homologous Series
- Paper II - Data handling of complex GC-MS signals to characterize homologous series as organic source tracers in atmospheric aerosols
- Paper III - Data handling of complex GC-MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source
- Paper IV - GC-MS analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosol: comparison between silylation and esterification derivatization procedures
- Paper V - Distribution of n-alkanes in the northern Italy aerosols: data handling of GC-MS signals for homologous series characterization

1

Introduction

Contents

1.1 Complex Mixtures: Multicomponent Chromatograms 3
1.2 Signal Processing 4

The aim of this Ph.D. project was a study of multicomponent chromatograms of complex mixtures, using a chemometric approach and to further develop it.

The activity has been concentrated in the study of analytical-separative methods (in particular Gas Cromatography-Mass Spectrometry, GC-MS) for complex samples of environmental interest, especially for PM (particulate matter) samples.

A fundamental part of this Ph.D. project has been dedicated to the development of mathematical and statistical algorithms for the data treatment of the GC-MS signal obtained from the analysis, in order to extract relevant information from the complex chromatogram, such as important indexes involved in the environmental studies.

In particular, the project involved the identification and the characterization of homologous series of organic compounds (n-alkanes and carboxylic acids) that could be usually found in environmental samples, because they contain fundamental information to distinguish, for example, different types of emission sources, anthropic or biogenic.

It has been developed a chemometric approach, which uses the AutoCoVariance Function (ACVF) computed on the digitized chromatogram, in order to quantificate the number of terms of the homologous series (n_{max}) and their distribution, with particular attention to the relative abundance and, consequently, the prevalance of the odd to even terms of the series (CPI).

This is one of the most important parameters (environmental biomarkers) to perform a study of source apportionment.

The method has been validated using simulated chromatograms and its applicability has been tested, with successful results, on real samples of known origin (e.g. gasoil or plant samples) and, finally, to particulate matter samples, obtained thanks to a collaboration with the research group of Environmental Sciences Department of the University of Milano Bicocca.

1.1 Complex Mixtures: Multicomponent Chromatograms

Environmental or natural samples often contain hundreds of components. A gasoil or crude oil products sample, for instance, contains tens of thousand of different components. Gas Chromatography-Mass Spectrometry (GC-MS) is a powerful tool to qualitatively and quantitatively analyze the composition of mixtures, especially the practical complex samples. GC-MS is being extensively used in scientific research and practical applications but an incomplete GC separation is a common problem.

Routine one-dimensional (1D) chromatographic methods cannot handle a complete qualitative and quantitative analysis of complex mixtures. For example, a 450-m long open tubular GC column with more than 1.3 million effective plates identified 970 compounds in a gasoline standard, yet that separation still had many unresolved peaks [1].

The complete chromatographic resolution of such complex samples requires tens of millions of theoretical plates, considering that peak capacity is roughly proportional to the square root of the number of theoretical plates. Because peaks do not elute equidistantly, analyzing a complex sample creates many coeluting components even with an extremely high-efficiency separation system.

The overlapping chromatographic peaks, which are generated by the incomplete separation in column, are often observed in practical analysis, since the compounds in such systems are complicated and often similar in properties.

These overlapping peaks may affect the qualitative analysis by mass spectral information and worsen the quantitative measurement by chromatographic peak area, and sometimes even make the analysis completely impossible.

The severe peak overlap often observed in such multicomponent separations arises mostly because of the random distribution of retention times and the limited peak capacity of the separation system.

To date GC analysis is a very rich source of data for chemical analysis, but extracting relevant information from the large, complex data sets is a challenge for information technologies. This is particularly true for hyphenated (GC-MS) and multidimensional (LC-GC, GC-GC, GC/GC) GC techniques, which generate data sets that are 2 or 3 orders of magnitude larger than for conventional GC [2, 3].

The quantity and complexity of GC data makes human analysis of GC signals difficult and time-consuming and motivates the need for computer-assisted signal processing to transform GC data into usable information. Advanced information technologies offer powerful solutions for many of the problems associated with the GC analysis: data handling, processing, analysis, and reporting. In particular, a mathematical approach is very useful to deconvolve incompletely resolved peaks and to interpret the chromatogram, extracting all the analytical information hidden therein, in other words decoding the complex chromatogram [4].

In a deconvolution process, a short section of a chromatogram, usually one cluster of several overlapping peaks, is investigated, and the profiles of the individual Single Component (SC) peaks are estimated with an algorithm. However, by using a statistical analysis, no specific information on a particular component is obtained, and the presence or absence of a compound cannot be determined, nor can its concentration be estimated.

The result is that the total chromatogram is regarded as a statistical ensemble whose common attributes, such as peak width, peak shape, extent of separation, number of detectable components, saturation of the separation space, and order/disorder of the peaks, are estimated [5, 6].

1.2 Signal Processing

When a sample has many compounds, different interval between adjacent peaks, peak clusters and void spaces are present in the multicomponent chromatogram. In other words, the retention patterns of complex mixtures can be remarkably different. This is because the distribution of the standard free energy differences between the stationary and mobile phases define a pseudorandom retention time distribution [7]. Accordingly, a complex chromatogram looks like a random series of peaks. Felinger et al. have proposed a method to decode complex multicomponent chromatograms, using Fourier analysis [8, 9]. Fourier transformation has been widely used for processing signals of analytical instruments, because several calculations are simpler in the frequency domain than in the time domain. Some models of chromatography also offer simpler solution in the frequency domain.

Fourier analysis considers the chromatogram as a finite-lenght fraction of a stochastic process; it means that the chromatogram of a complex mixture can be handled as a random series of peaks, that are uncorrelated random variables. The power spectrum of such a multicomponent chromatogram is calculated as either the time or the ensamble average of the *random* chromatogram.

Models have been derived for the power spectrum of various multicomponent chromatograms [10–12]. Fourier analysis can be applied to either ordered or disordered chromtograms as well [13]. The varying peak width and the peak height dispersion are taken into account by Fourier analysis theory. By means of the power spectrum or the autocovariance function (ACVF) of the chromatrograms, the mean peak width and the retention pattern can also be determined. The power spectrum is the square of the absolute value of the Fourier transformed signal, the ACVF is better described in the chapter 2.

According to Wiener-Khinchin theorem, the power spectrum and ACVF form a *Fourier pair.* Thus, the ACVF and the power spectrum are identical tools to characterize multicomponent chromatograms. Accordingly, in many instances it's not necessary to calculate the Fourier transform or the power spectrum of a chromatogram, it's sufficient to analyze the ACVF [6].

2

AutoCoVariance Function

Contents

2.1 Theory **6**
2.1.1 Poissonian-Type Chromatogram 9
2.1.2 Ordered-Type Chromatogram 10
2.2 Homologous Series **12**
2.3 Linearization **14**
2.4 Odd/Even Prevalence **18**

2.1 Theory

The chemometric approach studies the AutoCoVariance Function ($ACVF_{tot}$) that can be directly computed from the experimental chromatogram acquired in digitized form. The Experimental $ACVF_{tot}$ ($EACVF_{tot}$) at the correlation time Δt is given by the following expression [14]:

$$EACVF_{tot}(\Delta t) = \frac{1}{N_p} \sum_{j=1}^{N_p - s} (Y_j - \hat{Y}) \cdot (Y_{j+s} - \hat{Y}) \tag{2.1}$$

$$s = 0, 1, 2 ... M - 1$$

where Y_j is the digitized chromatogram signal, $\hat{Y}$ its mean value, N_p the number of points of the digitized chromatogram, and M the truncation point in the $EACVF_{tot}$ computation. The correlation time $\Delta t = s\tau$, where τ is the time interval between the subsequent digitized positions, and assumes discrete values with s ranging from 0 to M. $EACVF_{tot}$ represents the correlations between subsequent peaks in the chromatogram.

Theoretical expressions (theoretical ACVF, TACVF) have been developed to express ACVF in terms of the hidden separation parameters, i.e., number of SCs, m_{tot}, SC peak standard deviation, σ, the distribution of the SC retention pattern (Interdistance Model, IM) and abundance (Abundance Model, AM) [10]. They require theoretical models to describe complex chromatograms: many functions can be developed to describe the infinity of real cases. There are two limit cases of retention patterns: a Poissonian (P) distribution that describes a completely disordered separation where SC retention positions are uniform randomly distributed over the chromatographic axis, and an ordered (O) distribution [8, 10, 12, 15].

The simplest approach assumes chromatographic peaks of Gaussian shape with constant width, i.e. constant standard deviation σ: this assumption is usually true under optimized programmed temperature conditions.

The original complete procedure is based on the fitting of EACVF to TACVF to obtain information on sample complexity, for example the number of components m, and on the separation system, mean peak width, σ [8, 10, 12, 14–16]. A simplified procedure based on simple computation on EACVF and graphical inspection of the EACVF plot has also been developed to obtain the same information [6, 13, 17–23]. The autocorrelation function (EACF), representing the EACVF normalized to the value computed at $\Delta t = 0$, is more frequently used than the EACVF itself. EACF describes short- and long-range correlation between subsequent peak positions. When EACF is plotted VS retention interdistance (b), two informative regions are obtained in the EACF plot [4, 14, 16, 24]:

- the first part of the EACF contains the shortest-range correlation and depends only on the shape of the single-component peaks. It resembles the descending half of a Gaussian peak describing mean peak shape averaged on all the peaks in the chromatogram. Theoretical expressions have been derived [8, 10, 12, 15] and a simplified procedure has been developed [13] to extract information from this part of EACVF by a simple graphical inspection i.e., the number of components, m_{tot}, and the average peak width σ;

- the second part (i.e., widest-range correlation interdistance) is determined by the retention pattern: if peak positions are randomly distributed throughout the chromatogram, EACF assumes a value of nearly 0; if ordered positions are repeated, some positive peaks are

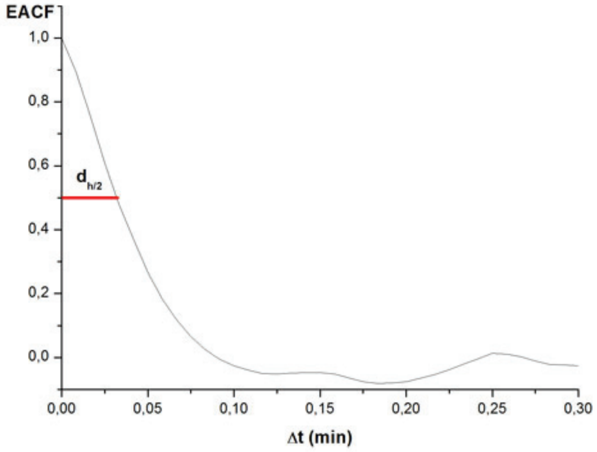


Figure 2.1: First region of EACF plot

present at the corresponding distance values in the EACF plot. This is the *deterministic* part of EACVF, resulting from a non-random retention pattern [4, 14, 16, 24].

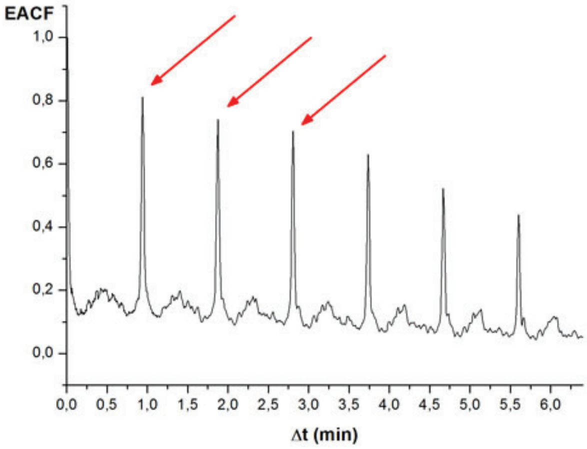


Figure 2.2: EACF plot with deterministic peaks (indicated by arrows)

2.1.1 Poissonian-Type Chromatogram

A disordered multicomponent chromatogram has m_P SCs (the suffix P refers to Poissonian retention pattern). The corresponding TACVF is given by the following equation [8]

$$TACVF_P(\Delta t) = \frac{A_{T,P}^2(\sigma_{h,P}^2/a_{h,P}^2+1)}{2\sqrt{\pi}m_P\sigma X}e^{-[(\Delta t)^2/4\sigma^2]} \tag{2.2}$$

where

$$A_{T,P} = \sqrt{2\pi}m_P a_{h,P}\sigma \tag{2.3}$$

is the total area of the chromatogram, $a_{h,P}$ and $\sigma_{h,P}^2$ are, respectively, the mean and the variance of SC peak abundance (i.e., peak maximum height of a SC peak), and X is the total time range of the chromatogram [8].

In the case of Poissonian retention pattern, the first region in the $EACVF_P$ plot ($0 < \Delta t \leq 4\sigma$), according to equation 2.2, is expected to be half of a Gaussian peak of standard deviation equal to $(2\sigma)^{1/2}$, showing a shape averaged on the shape of all the peaks present in the chromatogram. From $EACVF_P$ half-height peak width, $d_{h/2}$, the mean peak standard deviation can be simply estimated [16]:

$$\sigma = d_{h/2}/1.665 \tag{2.4}$$

From the value of $EACVF_P$ at the origin ($\Delta t = 0$), it's possible to estimate the number of SCs of the chromatogram, m_P, by rearranging equation 2.2 [16]:

$$m_P = \frac{A_{T,P}^2(\sigma_{h,P}^2/a_{h,P}^2+1)}{EACVF_P(0)d_{h/2}2.129X} \tag{2.5}$$

In this equation, the quantities $A_{T,P}$ and X can be determined from the experimental chromatogram and $d_{h/2}$ can be determined over the experimental $EACVF_P$ plot (equation 2.4). On the contrary, the exact value of the quantity $\sigma_{h,P}^2/a_{h,P}^2$ cannot be experimentally determined from the chromatogram due to SC peak overlapping. However, $\sigma_{h,P}^2/a_{h,P}^2$ can be approximated by the $\sigma_{M,P}^2/a_{M,P}^2$ value, which is the peak maximum dispersion ratio, i.e., the dispersion ratio computed from the observed peak maximums in the chromatogram:

$$\sigma_{h,P}^2/a_{h,P}^2 \approx \sigma_{M,P}^2/a_{M,P}^2 \tag{2.6}$$

Consequently m_P can be estimated by using the following equation:

$$m_P = \frac{A_{T,P}^2(\sigma_{M,P}^2/a_{M,P}^2+1)}{EACVF_P(0)d_{h/2}2.129X} \tag{2.7}$$

The quantity m_P is related to the Poisson distribution, and its standard deviation is known to be equal to $m_P^{1/2}$. Consequently, equation 2.7 yields an estimate $m_{est,P} = m_P \pm m_P^{1/2}$.
Equations 2.6 and 2.7 can be used not only in the case of a strictly Poissonian retention pattern but also to other cases of random patterns [5,8,10,12–16]. A useful form of equation 2.7 can be obtained if the total area of the chromatogram $A_{T,P}$ is expressed as

$$A_{T,P} = m_P A_{m_P,P} \tag{2.8}$$

where $A_{m_P,P}$ is the average area of SC peaks in the random multicomponent chromatogram. By combining equations 2.7 and 2.8, it's possible to obtain

$$m_P = \frac{EACVF_P(0) d_{h/2} 2.129 X}{A^2_{m_P,P}(\sigma^2_{M,P}/a^2_{M,P} + 1)} \tag{2.9}$$

Equation 2.9 clearly shows the direct proportionality between m_P and $EACVF_P(0)$ under conditions of constancy of $A_{m,P}$ value [4].

2.1.2 Ordered-Type Chromatogram

An ordered pattern in a multicomponent chromatogram (i.e. an homologous series) is formed by a sequence of n_{max} SC peaks where the retention time (t_R) of the n^th term is described by:

$$t_R(n) = c + bn \tag{2.10}$$

$$n = 0, 1, 2, 3 ... n_{max}$$

where c represents the contribution of a specific functional group to the overall retention, and b is the retention increment between terms of the homologous series, e.g. the CH_2 retention time increment, in the strict case of GC analysis under optimized linearized temperature programming conditions [4]

If this condition is not met in practice, a linearization algorithm can be applied to rescale the original signal in order to obtain the same peak width (σ values) and constant retention increment between subsequent terms of the series.

If the term with $n = 0$ is absent in the considered homologous series, the number of SCs will be equal to n_{max}. The constants c and b are called the phase and frequency indicators of the sequence, respectively, according to Giddings definition [25]. In this case, the expression of

TACVF can be obtained from ref [26], under the assumption that the SC term with $n = 0$ (see equation 2.10) is absent:

$$TACVF_O(\Delta t) = \sum_{k=0}^{k=n_{max}-1} \frac{A_{T,O}^2}{2\sqrt{\pi}\sigma X(n_{max}-k)} \left(\frac{\sigma_{h,O}^2}{a_{h,O}^2}+1\right) e^{-[(\Delta t-bk)^2/4\sigma^2]} \tag{2.11}$$

$$A_{T,O} = \sqrt{2\pi} n_{max} a_{h,O} \sigma \tag{2.12}$$

where $A_{T,O}$ is the total area of the ordered chromatogram, $a_{h,O}$ and $\sigma_{h,O}^2$ are respectively the mean and the variance of SC peak abundance (i.e., peak maximum height of a SC peak) in the O-type chromatogram. Note that this expression does not contain the phase c, and thus, information concerning the sequence phase is lost: the ACVF retains only the recursivity of an ordered structure. According to equation 2.11, the $TACVF_O$, and therefore the $EACVF_O$ plot, shows well-defined Gaussian peaks of standard deviation equal to $(2\sigma)^{1/2}$, located at interdistances bk, corresponding to repeated interdistances between terms of the homologous series (equation 2.11). These peaks are called deterministic since they reflect the order of the sequence; their height decreases on k, but their shape is independent of k. The subscript O identifies that an ordered retention pattern is present in the multicomponent chromatogram.

Equations similar to equations 2.7 and 2.9 can be derived even in the case of O multicomponent chromatograms. Starting from the same assumptions employed before in deriving those equations, it's possible to write:

$$n_{max} - k = \frac{A_{T,O}^2(\sigma_{M,O}^2/a_{M,O}^2+1)}{EACVF_O(bk)d_{h/2}2.129X} \tag{2.13}$$

$$n_{max} - k = \frac{EACVF_O(bk)d_{h/2}2.129X}{A_{n_{max},O}^2(\sigma_{M,O}^2/a_{M,O}^2+1)} \tag{2.14}$$

where the quantity $A_{n_{max},O}$, is the mean area of SC peaks defined by

$$A_{T,O} = n_{max} A_{n_{max},O} \tag{2.15}$$

and $\sigma_{M,O}^2/a_{M,O}^2$ is the peak maximum dispersion ratio. The different peaks in the $EACVF_O$ plot will be well distinct provided that $b > 4\sigma$ (see equation 2.11). In this case, the SC number n_{max} can be determined by using the peak at the origin, in analogy with equation 2.9:

$$n_{max} = \frac{A_{T,O}^2(\sigma_{M,O}^2/a_{M,O}^2+1)}{EACVF_O(0)d_{h/2}2.129X} \tag{2.16}$$

The first deterministic peak can also be used and in this case

$$n_{max} = \frac{A^2_{T,O}(\sigma^2_{M,O}/a^2_{M,O} + 1)}{EACVF_O(b)d_{h/2}2.129X} \tag{2.17}$$

Consequently, an experimental check of the ordered character of the chromatogram can be obtained by comparing the n_{max} values estimated from equations 2.16 and 2.17. $A_{T,O}$, X, and $d_{1/2}$ values are experimentally accessible parameters [4].

2.2 Homologous Series

A complex mixture, and the relative multicomponent chromatogram, may be formed by one homologous series of n_{max} SCs (ordered sequence, equation 2.10) and of an ensemble of uncorrelated m_P SCs, (random component), the total number of SCs, m_{tot}, and the total area of the chromatogram will be given by

$$m_{tot} = m_P + n_{max} \tag{2.18}$$

and

$$A_{T,tot} = A_{T,P} + A_{T,O} \tag{2.19}$$

respectively. It has been developed a method to estimate m_{tot}, m_P, b, and n_{max} from the $EACVF_{tot}$ plot of the chromatogram, under specific conditions. In analogy with equations 2.7 and 2.9:

$$m_{tot} = \frac{A^2_{T,tot}(\sigma^2_{M,tot}/a^2_{M,tot} + 1)}{EACVF_{tot}(0)d_{h/2}2.129X} \tag{2.20}$$

$$m_{tot} = \frac{EACVF_{tot}(0)d_{h/2}2.129X}{A^2_{m_{tot},tot}(\sigma^2_{M,tot}/a^2_{M,tot} + 1)} \tag{2.21}$$

where the quantity $A_{m_{tot},tot}$ is the mean area of SC peaks defined by

$$A_{T,tot} = m_{tot}A_{m_{tot},O} \tag{2.22}$$

and $\sigma^2_{M,tot}/a^2_{M,tot}$ is the peak maximum dispersion ratio in the chromatogram. Equation 2.21 can be obtained by combining equation 2.14 with $k = 0$, equations 2.9 and 2.18, and further by assuming that

$$EACVF_{tot}(0) = EACVF_P(0) + EACVF_O(0) \tag{2.23}$$

$$A_{m_{tot},tot} \approx A_{m_P,P} \approx A_{n_{max},O} \tag{2.24}$$

$$\sigma^2_{M,tot}/a^2_{M,tot} \approx \sigma^2_{M,P}/a^2_{M,P} \approx \sigma^2_{M,O}/a^2_{M,O} \tag{2.25}$$

Equations 2.24 and 2.25 mean that SCs in both the ordered and the Poissonian components of the multicomponent chromatogram have equal average peak areas and peak height dispersion ratios, respectively. Equation 2.36 expresses the rule of the variance additivity for independent variables, remembering that $EACVF(0)$ has the meaning of a variance (see equation 2.1). This condition holds true for the present case since the Poissonian part of the chromatogram is completely random and thus not correlated with the ordered one. Consequently, m_{tot} can be estimated from $EACVF_{tot}$ by using equation 2.20, under the conditions described by equations 2.24 and 2.25.

The n_{max} value, and therefore also the m_P value, can be obtained from $EACVF_{tot}$ under given conditions. In fact, assuming that, for $k > 0$

$$EACVF_{tot}(bk) = EACVF_O(bk) \tag{2.26}$$

equation 2.14, together with the conditions expressed by equations 2.36,2.24,2.25 and 2.26 becomes:

$$n_{max} - k = \frac{EACVF_{tot}(bk)d_{h/2}2.129X}{A^2_{m_{tot},tot}(\sigma^2_{M,tot}/a^2_{M,tot}+1)} \tag{2.27}$$

$$b \geq 4\sigma$$

Equation 2.27 means that that n_{max} can be evaluated from $EACVF_{tot}$, even if under strict conditions. The condition $b \geq 4\sigma$ means that the SC peaks belonging to the homologous series are each other sufficiently resolved in the chromatogram (equation 2.10) and their correlation does not interfere with that inside a SC component peak. In this case, the first and the subsequent deterministic peaks of the $EACVF_O$, for example the $EACVF_O(bk)$ peaks for $k > 0$, are well separated from the origin ($\Delta t = 0$), beyond $\Delta t = 4\sigma$. Under these conditions, equation 2.26 also holds true since the $EACVF_P \approx 0$, in the region $\Delta t \geq 4\sigma$, and the $EACVF_{tot}$ plot is only made by the $EACVF_O$.

A simple form of equation 2.27 can be obtained for $k = 1$, by combining equation 2.21 and equation 2.27:

$$n_{max} = m_{tot}\frac{EACVF_{tot}(b)}{EACVF_{tot}(0)} + 1 \tag{2.28}$$

$$b \geq 4\sigma$$

The conditions expressed by equations 2.24 and 2.25 require a comment since they are not very common in the practice. In particular, the hypothesis expressed by equation 2.24, that means the average SC abundance of the SCs belonging to the total mixture and that to a given homologous series are the same, seems critical: in fact, it is often possible the condition that one homologous series can be either predominant or in trace respect to the majority of the other SCs. Otherwise, equation 2.25 can be more or less met in practice since it establishes that the degree of randomness on the SC peak height dispersion ratio is similar for either the random component or the ordered component (homologous series).

It is possible, moreover, that the equation 2.24 is not strictly true. By combining equations 2.20 or 2.21 with equations 2.13 or 2.14, respectively, for $k = 1$, $b \geq 4\sigma$ and under the conditions described by equations 2.36,2.24,2.25 and 2.26, it's possible to obtain:

$$n_{max} = m_{tot} \frac{EACVF_{tot}(0)}{EACVF_{tot}(b)} \cdot \frac{A^2_{T,O}}{A^2_{T,tot}} + 1 \qquad (2.29)$$

$$b \geq 4\sigma$$

and

$$n_{max} = m_{tot} \frac{EACVF_{tot}(b)}{EACVF_{tot}(0)} \cdot \frac{A^2_{m_{tot},tot}}{A^2_{n_{max},O}} + 1 \qquad (2.30)$$

$$b \geq 4\sigma$$

respectively. Consequently, in the general case, when the ratio of the SC area of structured class, with respect to that of the totality of the SCs is unknown, the sole $EACVF$ ratio and m_{tot} (obtained from equation 2.20) cannot yield a quantitative estimate of SC number of the homologous series but only an *apparent* SC number.

Equations 2.29 and 2.30 suggest that the use of selective detectors combined to universal ones, could be a useful strategy of study. In fact they make it possible to selectively detect specific compound classes, and thus, the contribution of the different classes can be decoded from the total mixture and quantitatively estimated [4].

2.3 Linearization

The study of $EACVF_{tot}$ to identify the sample chemical composition and extract structural information regarding the mixture components from the GC signal is based on a strict linear relation between the retention time $t_R(n)$ and number of repeated units n within a homologous series (Eq. 2.10). This is true under linear temperature-programmed GC conditions, as confirmed by both experimental evidence and theoretical studies based on retention thermodynamics [20, 27, 28].

However, the strictly homogenous retention pattern yielding constant retention increments between subsequent terms of homologous series is difficult to be achieved in the practice because of experimental limitations, i.e., the not strictly linear temperature-programmed GC runs, poor reproducibility in flow rate or temperature, variations in injection-timing and temperature program rate [27, 28].

Therefore, in order to usefully apply the $EACVF_{tot}$ procedure, a data handling algorithm is required to linearize experimental chromatograms prior to $EACVF_{tot}$ computation.

If $Y(x)$ represents the chromatographic signal, where x is the retention time, the time axis is transformed into a new scale by using a function $z = g(x)$ to relate the original time axis to the new z axis. To preserve the total signal area, the following condition must be fulfilled:

$$Y_1(x)dx = Y_2(z)dz \qquad (2.31)$$

Instead of a continuous function $z = g(x)$, it's possible to use an empirical transformation procedure based on an equidistant retention position between the subsequent terms of homologous series Δx (for example the addition of a CH_2 group in a n-alkane homologous series). This means that the applied transformation has the property:

$$Y_1(x)\Delta x = Y_2(z)\Delta z \qquad (2.32)$$

The use of certified standard homologous series compounds as external reference to build up a GC retention scale is very common in Gas-Chromatography: in fact, the standards (for example n-alkanes series) may act as the flexible *mile-stone* system of the chromatogram, and the relative position of the analyte compounds can be referred to them [29, 30].

A first procedure has been developed [20, 24, 31] and it worked, through a Fortran77 algorithm, in this way: a homologous series reference mixture containing the terms displaying retention values in the same range as the sample was analyzed under the same temperature-programmed GC conditions used for the unknown sample. Within a given threshold distance, the reference peaks are matched to the nearest peaks in the sample chromatogram. The sample signal was divided into many regions corresponding to the distance $\Delta x = t_R$ between subsequent terms of the series; a Δx value (usually the average of experimental Δx values) was selected as constant Δz retention increment in the new scale. Each inter-peak region is taken individually and it is stretched, or shrunk, to force each Δx interdistance to the constant Δz value.

Recently a new, but similar, procedure has been developed; it's based on the same property (equation 2.32). A MATLAB® algorithm was generated starting from the Fortran77 one, in order to quickly and more correctly linearize the chromatogram (see Appendix 6 A for the complete algorithm). After the recognition of the homologous series terms in the unknown sample chromatogram (by a simple comparison between standard references and unknow peaks t_R), the maximum Δx_i was chosen by the program itself and all the homologous series terms interdistance (Δx_i) were stretched to reach the $MAX(\Delta x_i)$ value. The stretching step uses a interpolation function in order to preserve the total area of the chromatogram.

If terms of an homologous series are present in the sample, a structured chromatogram characterized by retention repetition is obtained after the linearization procedure. The order introduced into the signal by this tool can be simply singled out by the experimental autocorrelation function ($EACF_{tot}$) plot [20, 31]. Figure 2.3 shows a standard mixture containing six subsequent terms of a n-alkane series ($C_7 - C_{12}$) in addition to 14 organic compounds with uncorrelated structures. The analysis was performed with this temperature program: $30°C$ for $3min$, an increase to $80°C$ at $5°C/min$. Figure 2.3(a) shows the original chromatograms of the standard mixture and the reference $C_7 - C_{12}$ n-alkanes (in the inset) and figure 2.3(b) shows the linearized chromatograms of the standard mixture and the reference in the inset: the arrows indicate the $\Delta t_R = 2.5min$ value selected as the constant retention increment [31].

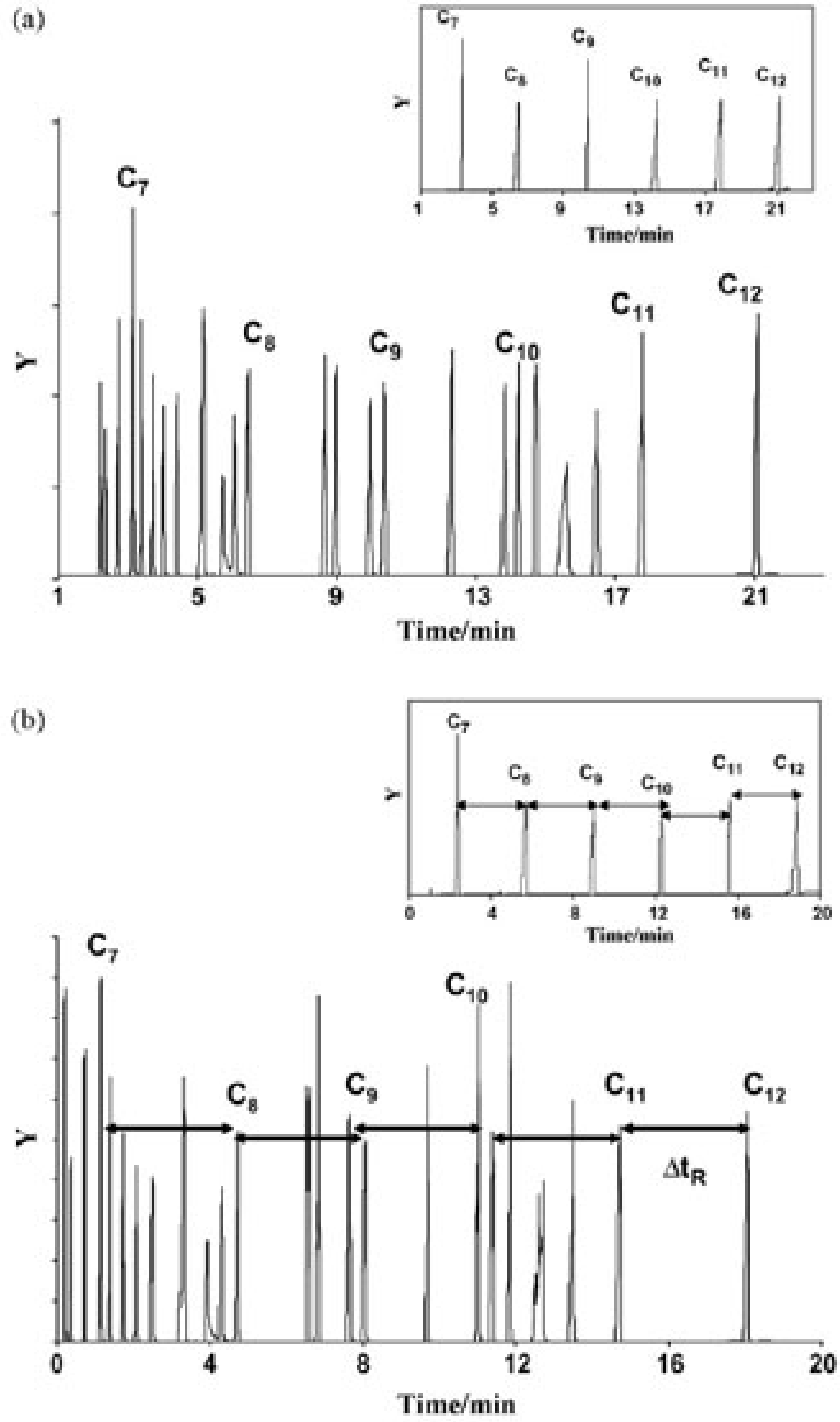


Figure 2.3: GC-MS chromatogram of a not linearized mixture and the relative linearized one [31]

2.4 Odd/Even Prevalence

The EACVF method has been further developed to extract information about the abundance distribution of the terms of an ordered pattern in a multicomponent chromatogram, like a homologous series, present in a complex mixture. A MATLAB® algorithm has been developed to estimate a useful index (R) of this abundance distribution (see Appendix 6 B for the complete algorithm) [32].

In order to build the model, the chromatogram of the homologous series (equation 2.10) is described as the combination of two sequences of peaks representing odd and even terms of the homologous series; they are located at a double repeated interdistance $\Delta t = 2b$ shifted by the quantity b (odd (o) and even (e) sequences containing n_o and n_e peaks, respectively, see Figure 2.4(a).

In this simplified approach the series contains the same number of odd and even terms $n_o = n_e = n$ to yield a total number of terms of the series $n_{max} = 2n$. The $EACVF$ is computed on the signal of each sequence ($EACVF_o$ and $EACVF_e$): their plots show deterministic peaks at the constant interdistance values $\Delta t = 2bk$ according to the following equations (see Figure 2.4(b)):

$$EACVF_o(2bk) = \frac{\sqrt{\pi}\sigma a_{o,h}^2}{X}\left[\frac{\sigma_{o,h}^2}{a_{o,h}^2} + 1\right](n_o - k) \tag{2.33}$$

$$EACVF_e(2bk) = \frac{\sqrt{\pi}\sigma a_{e,h}^2}{X}\left[\frac{\sigma_{e,h}^2}{a_{e,h}^2} + 1\right](n_e - k) \tag{2.34}$$

where $a_{o,h}$, $a_{e,h}$ and $\sigma_{o,h}^2$, $\sigma_{e,h}^2$ are the mean value and the variance of the SC peak heights of the odd and even sequences, respectively.

The whole series containing $n_{max} = 2n$ terms is obtained by superimposing the o and e sequences and the $EACVF_{tot}$ computed on the total chromatogram can be investigated as a combination of $EACVF_o$ and $EACVF_e$. To handle the $EACVF_{tot}$, new equations are derived in order to extract information on the odd/even prevalence of the sequence terms (see the Appendix of [32].

It is assumed that both the odd and even terms display the same peak abundance distribution, described by peak height dispersion ratio σ_h^2/a_h^2:

$$\frac{\sigma_{o,h}^2}{a_{o,h}^2} = \frac{\sigma_{e,h}^2}{a_{e,h}^2} = \frac{\sigma_h^2}{a_h^2} \tag{2.35}$$

This condition is usually met in real samples since the compound abundances generally follow the most probable Exponential distribution, yielding: $\sigma^2/a^2 = 1$.

It can be demonstrated that $EACVF_{tot}(bk)$ values at $\Delta t = bk$ for even k terms are obtained by combining $EACVF_o(bk)$ and $EACVF_e(bk)$ values to yield an equation related to the addition of the two series abundances [32]:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma(a_{o,h}^2 + a_{e,h}^2)(n-k)}{X}\left[\frac{\sigma_h^2}{a_h^2}+1\right] \tag{2.36}$$

$$k = 0, 2, 4, ...2n-2$$

At $\Delta t = bk$ for odd k values, the $EACVF_{tot}(bk)$ values are given by the cross-correlation term between components of the o and e sequences:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma 2(a_{o,h}\cdot a_{e,h})(n-k)}{X}\left[\frac{\sigma_h^2}{a_h^2}+1\right] \tag{2.37}$$

$$k = 1, 3, 5, ...2n-1$$

Therefore, the $EACVF_{tot}(bk)$ peaks computed at subsequent k values give information on the specific abundance distribution pattern of the odd/even terms of the homologous series (Figure 2.4(b)). In fact, if the odd and even terms display the same mean abundance distribution ($a_{o,h} \approx a_{e,h}$) equation 2.36 and 2.37 are identical and the $EACVF_{tot}(bk)$ values are proportional to the values of the sequence $(2n-k)$ for $k = 1, 3, ...(2n-1)$. Any deviation from such a pattern is diagnostic of the presence of odd/even prevalence among the terms of the series.

To describe a specific odd/even distribution pattern for the terms of the homologous series, the R value is defined as the ratio between the mean value of the SC peak height of odd vs. even terms:

$$R = \frac{a_{o,h}}{a_{e,h}} \tag{2.38}$$

By computing equation 2.36 and 2.37 for $k = 2$ and $k = 1$, respectively, and introducing the R parameter, it's possible to write:

$$EACVF_{tot}(2b) = \frac{\sqrt{\pi}\sigma}{X}(a_{o,h}^2)\left(1+\frac{1}{R^2}\right)\left[\frac{\sigma_h^2}{a_h^2}+1\right](n-2) \tag{2.39}$$

$$EACVF_{tot}(b) = \frac{\sqrt{\pi}\sigma}{X}2a_{o,h}^2\frac{1}{R}\left[\frac{\sigma_h^2}{a_h^2}+1\right](n-1) \tag{2.40}$$

By dividing equation 2.40 by equation 2.40, the following expression can be obtained as a function of R:

$$\frac{EACVF_{tot}(b)}{EACVF_{tot}(2b)} = \frac{\frac{2}{R}(n_{max} - 1}{\left(1 + \frac{1}{R^2}\right)(n_{max} - 2} = \frac{2R(n_{max} - 1)}{(R^2 + 1)(n_{max} - 2)} \tag{2.41}$$

The equation can be simplified by introducing the approximation that the ratio between $(n_{max} - 1)$ and $(n_{max} - 2)$ is equal to 1: this is strictly true for large n_{max} values $(n \to \infty)$, otherwise it can be applied once n_{max} is known.

With this assumption equation 2.41 can be simplified into:

$$\frac{EACVF_{tot}(b)}{EACVF_{tot}(2b)} == \frac{2R}{(R^2 + 1)} \tag{2.42}$$

This is a simple quadratic equation, that can be solved to obtain the R value, considering Y as the ratio between the two $EACVF_{tot}$ values:

$$R = \frac{2 \pm \sqrt{4 - 4Y^2}}{2Y} \tag{2.43}$$

Equation 2.42 shows that the odd/even prevalence of the terms of the homologous series, expressed by the R value, can be directly estimated from the whole chromatogram by computing the $EACVF_{tot}$ values at $\Delta t = b$ and $\Delta t = 2b$ on the total signal [23,32]. Figure 2.4 shows a simulated chromatogram of a mixture formed by 5 odd and 5 even terms of a sequence, displaying abundance values generated according to an Exponential AM: mean abundance distributions were simulated for odd and even terms, to yield $R = 2$. Each series is formed by 5 terms located at a repeated interdistance $\Delta t = 2b = 3.60min$ shifted by a quantity $\Delta t = b = 1.80min$.

Figure 2.4(a)is a PC-generated chromatographic signal; figure 2.4(b) is the $EACVF_{tot}$ plot computed on the signal: the $EACVF_{tot}$ peaks diagnostic of the sequence at $\Delta t = b = 1.80min$ and $\Delta t = 2b = 3.60min$ are identified by the arrows [32]. If the general model based on the following equation

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma a_h^2(n_{max} - k)}{X}\left[\frac{\sigma_h^2}{a_h^2} + 1\right] \tag{2.44}$$

$$k=0,1,2,3,....n_{max}\text{-}1$$

is applied to estimate n_{max}, it may yield misleading results in the case of an odd/even prevalence: in fact, at $\Delta t = bk$ for odd k values, the $EACVF_{tot}$ values strongly depend on the presence of the odd/even prevalence in the peak abundance since it is related to the product $(a_{o,h} \cdot a_{e,h})$

(equation 2.37)). Otherwise, at $\Delta t = bk$ for even k values, the $EACVF_{tot}$ values are independent of the peak abundance distribution of the odd and even terms since it is related to the quantity $(a_{o,h}^2 + a_{e,h}^2)$ (equation 2.36): therefore, at $\Delta t = bk$ for even k values, $EACVF_{tot}$ values are used to obtain a correct estimation of n_{max}. In order to make the procedure more robust, for even k values, the computation is based on two subsequent $EACVF_{tot}$ deterministic peaks at $\Delta t = bk$ and $\Delta t = b(k+2)$ according to the following equation:

$$n_{max} = 2\frac{EACVF_{tot}(bk)}{EACVF_{tot}(b(k+2))} + k \tag{2.45}$$

The correct estimation of n_{max} based on equation 2.45 makes it possible to achieve an accurate estimation of R by using the rigorous equation 2.41 to remove the approximation introduced in equation 2.42. It must be underlined that the mathematic model developed on the basis of equations 2.41 and 2.45, strictly derived for a chromatogram that only contains homologous series terms, is applicable to the general case of complex mixtures containing random uncorrelated compounds in addition to the homologous series. In fact, the Poissonian component yields $EACVF_{tot}$ values significantly different from 0 only for $\Delta t \leq 4\sigma$, so that, at the repeated interdistances ($\Delta t = bk$), the $EACVF_{tot}$ values are mainly due to the contribution of the homologous series (equation 2.44) and can be used to evaluate its properties [23].

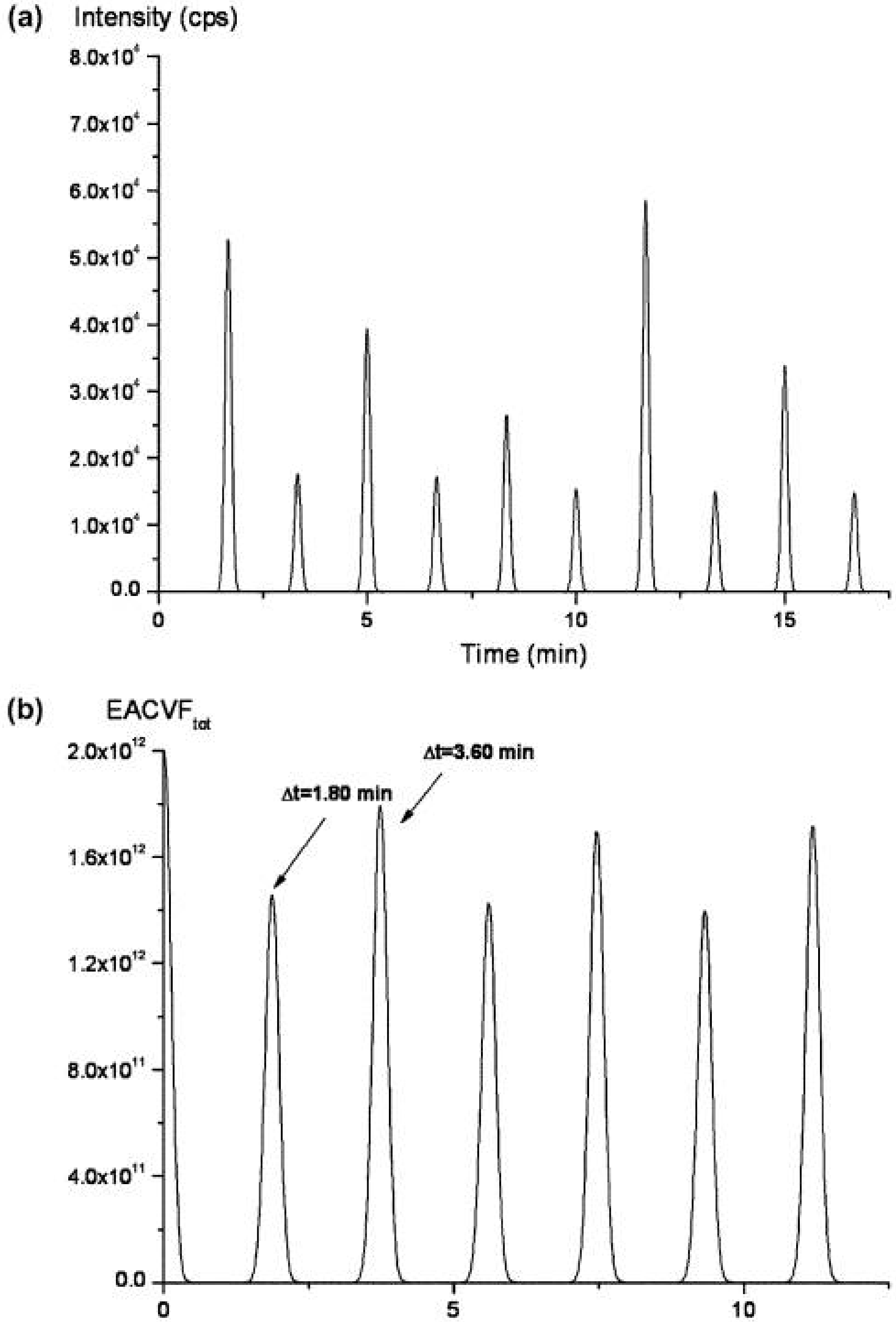


Figure 2.4: Simulated chromatogram and the relative $EACVF_{tot}$ with a R value equal to 2 [31]

3

Particulate Matter

Contents

3.1 Organic Fraction of PM 25

3.2 Markers and Source Apportionment 27

3.2.1 n-Alkanes 28

3.2.2 Carboxylic Acids 30

Particulate matter (PM) is a complex mixture which includes many different types of compounds. A comprehensive and correct characterization of the chemical composition of the PM is one of the most important steps in a pollution or, generally, in an environmental study. Because of its complexity, it's very difficult to separate, using traditional analytical methods, all the single components of a particulate matter sample. It is necessary to find out classes of compounds or specific components in order to use them as molecular tracers or biomarkers for a determined environmental study, such as a source apportionment study.

Air pollution associated to aerosols has recently gained the concern of scientific institutions and Public Agencies since an additional health risk for humans has been proven to arise from exposure to fine particles in general and specifically to their organic components [33].
Various epidemiological studies have shown associations between daily ambient concentrations of particulate matter and morbidity and mortality [34–36]. The observed effects became particularly clear when the mass concentrations of particles with aerodynamic diameter less than

$2.5 \mu m$ ($PM_{2.5}$, fine particles) were considered.

Apart from the particle number and/or mass concentration the chemical composition of particles may also be important for assessment of health effects, for example by determining the reaction pattern in the respiratory tract or the response of the whole body.

The results of epidemiological studies together with animal toxicology and in vitro experimental studies support the hypothesis that both physical (particle size, shape, surface) and chemical (dissolved and adsorbed chemicals, surface catalytic reactions) properties of the particles are involved in toxic, genotoxic and carcinogenic mechanisms of inhaled particulates [37].

It was recently recognised that inhaled ultra fine particles ($D_p < 100nm$) are more toxic than PM_{10} particles [38–40]. Their relative large surface area and the ability to be absorbed into tissues might be important factors in cardiopulmonary toxicity. However, the number of ultra fine particles in the air is often poorly correlated with $PM_{2.5}$ and even less with PM_{10}. Thus, ultra fine particles are unlikely to explain much of the association between particulate mass and health conditions. Also the impact of surface absorbed compounds on health outcomes is not well understood yet.

Epidemiological investigations of the influence of individual, particle bound chemical pollutants were done only for few inorganic species [41]. The role of transition metals (Fe, V, Zn) for acute reactions is under discussion [42]. However, little is known on the influence of the organic chemicals present in ambient particulate matter (PM) on the health outcomes. The influence of organic substances was evaluated by measurement of the concentration of elemental and organic carbon (EC/OC) [43]. But so far, the association of individual specific organic pollutants or groups of pollutants with health effects, occurring in the fine dust, was not examined in epidemiological studies.

For a time-series study, on the influence of organic aerosol compounds, it is necessary to have data of several compounds or groups of compounds at least with a daily resolution. Because most of the organic compounds occur in low concentrations in ambient aerosol, time-consuming analytical methods are required for their analysis.

Several studies address the organic composition of ambient PM, using gas chromatography-mass spectrometry (GC-MS) for separation and identification of semi volatile organic compounds (SVOC). GC-MS is a well established technique for the separation and analysis of complex mixtures [44, 45].

Investigations of the origin of air pollution, its major sources and the import of pollution from distant regions (Source Apportionment) become increasingly important due to the new standards which were set for particle mass less than $10\mu m$ (PM_{10}) in the EC-Directive 99/30/EC (EC Directive, 1999). Together with the Air Quality Framework Directive 96/92/EC (EC Directive, 1996) authorities in non-attainment zones are required to assess PM_{10} pollution [46].

3.1 Organic Fraction of PM

The atmosphere is a processing unit for organic compounds. They are ubiquitous and abundant in ambient aerosols. They typically account for 20-50% of the fine particle mass [47–51] and are often internally mixed in the same particles with inorganic aerosols [52–54]. Organic compounds play important roles in the formation, growth, and removal of ambient aerosols [55]. They also significantly affect the hygroscopicity [56], toxicity [57], direct radiative properties [58, 59], and indirect effects [60] of atmospheric aerosols and therefore have major implications for climate, visibility, and human health.

Elucidating the urban-to-global roles as well as the sources and fate of atmospheric aerosols inherently must rely on a thorough understanding of the chemical and microphysical properties of particulate organics. However, it is extremely difficult to obtain a complete description of the molecular composition of aerosol organics because of the number, complexity, and extreme range of physical and chemical properties of these compounds [61].

Organic compounds are important atmospheric components. The formation of organic aerosols represents one of the removal processes of volatile organic compounds (VOC). Thus, organic compounds play an important role in photochemical reactions leading to ozone formation [62]. On the other hand, they compete with inorganic compounds for oxidising species such as ozone, hydroxyl and nitrate radicals [63]. If organic aerosols occur in the submicron range they can originate cloud condensation nuclei [64]. Organic aerosols have been associated with indirect climate forcing, because they have optical properties and contribute to visibility degradation [65]. Organic aerosols may change chemical, optical and hygroscopic properties of inorganic aerosols [56]. The presence of some components (e.g. polyaromatic hydrocarbons) is a cause of concern since they have proven carcinogenic and/or mutagenic properties [66–68]. A large fraction of organics is associated with particles smaller than 3 mm, which can reach the respiratory system [69].

Secondary organic aerosols (SOA) are formed from both biogenic and anthropogenic gaseous precursors. The major biogenic compounds involved in aerosol formation are considered to be monoterpenes, which constitute more than 80% of the VOC emissions from conifers [69]. Approximately 50% of anthropogenic VOC are emitted by mobile sources, while industrial sources represent the second greatest VOC emitter [63]. The formation of SOA can follow complex chemical pathways, many of which remain unknown. Despite the uncertainties, it is recognised that organic aerosol formation is typically dominated by $C_5 - C_{10}$ species, because compounds with more than 10 carbons tend to be present at low concentration and species with small molecular weights have high saturation vapour pressures [63]. Organic species can form new aerosols by condensation or react heterogeneously on pre-existing aerosols [70]. Even when products are present at less than their saturation vapour pressure, they may still condense onto existing aerosols [71,72]. Species such as the hydroxyl radicals and ozone are expected to be atmospheric oxidants of hydrocarbons leading to products containing carbonyl (-C=O), carboxy (-COOH), and hydroxy (-OH) functional groups. OH radicals attack alkanes. C_4 initiating their oxidation. Alkoxy radical intermediates are formed, which through isomerisation leads to the formation of carbonyl products. For hydrocarbons containing double bonds (alkenes for example) hydroxyl radical or ozone can start oxidation. The next reactions form products such as carbonyls, hydroxy carbonyls, dicarbonyls, carboxylic acids, and oxocarboxylic acids [47,73].

A detailed understanding of SOA formation in the atmosphere is essential to characterise the chemical composition of ambient organic aerosols, to accurately incorporate such processes in air quality models, and to be able to attribute the ambient organic aerosol mass to the appropriate man-made and natural sources.

Long chain n-alkanes, n-alkanols, n-alkanals, 2-alkanones, n-alkanoic acids, n-alkanoic acids salts, $\alpha - \omega-$dicarboxylic acids, polycyclic aromatic hydrocarbons (PAHs), and their oxygenated and nitrated derivatives have been detected in urban and rural, as well as in remote marine aerosol [74].

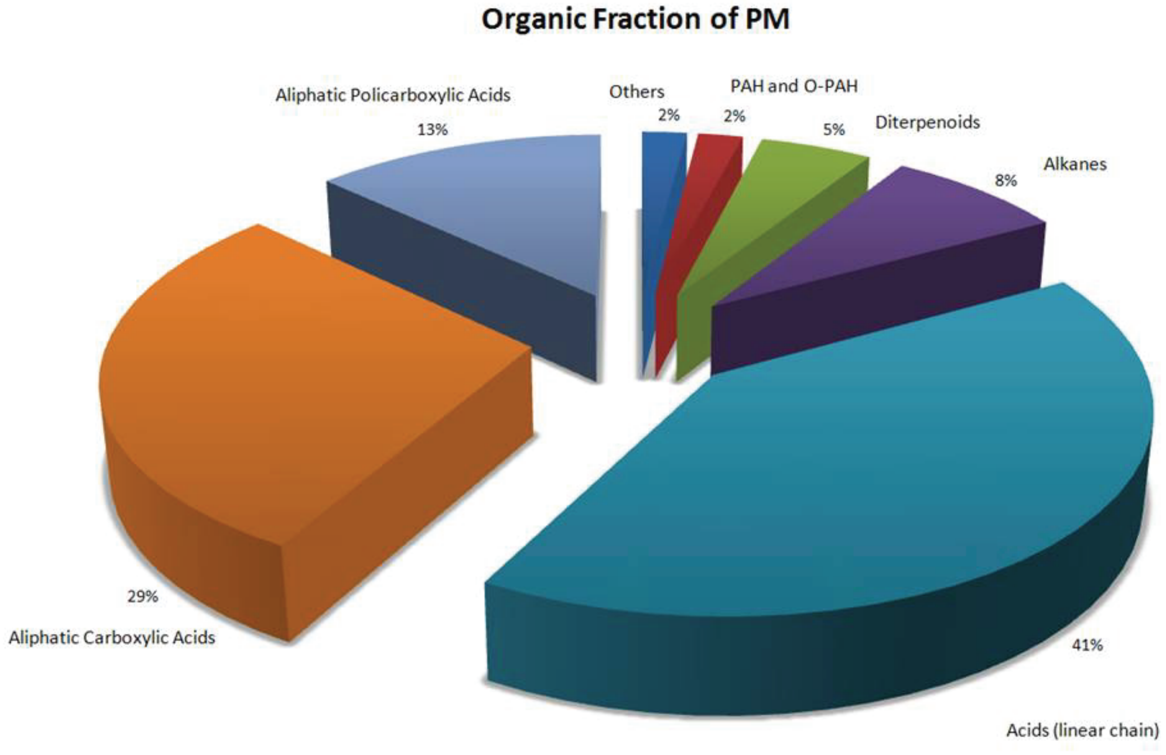


Figure 3.1: Organic fraction of Particulate Matter [75]

3.2 Markers and Source Apportionment

The organic chemical composition of airborne fine particles is an important, multidisciplinary research area for several reasons.

First, controlling fine-particle atmospheric concentrations requires an understanding of the emission sources. Organic complex mixtures contain molecular tracers that can be linked to specific emission sources or are byproducts from dominant atmospheric photochemical reactions. Identifying the mass contribution of key source markers in $PM_{2.5}$ complex mixtures and coupling this information with chemical mass balance (CMB) models, for example, provides a quantitative approach for estimating individual emission source inputs to urban atmospheres [75–77].

Second, full chemical descriptions of organic mixtures collected as $PM_{2.5}$ have not been achieved. Approximately 20% of the masses of organic complex mixtures are resolved quantitatively as individual compounds [75, 76, 78, 79]. The remaining 80% of this organic complex mixture mass may contain individual compounds with great significance as ambient indicators representing particulate matter exposure. Incomplete chemical descriptions of fine-particle complex organic mixtures have slowed progress in establishing critical links between specific toxic constituents of airborne particles with health indicators [80].

Two chemical classes of compounds have been mainly investigated in this thesis, due to their relevance in environmental chemistry as molecular tracers: n-alkanes (section 3.2.1) and carboxylic acids (section 3.2.2).

3.2.1 n-Alkanes

Identification and quantification of specific compounds as chemical markers is a convenient approach to characterize the samples formed by a complex mixture of organics. Extensive studies have demonstrated that n-alkanes are especially suited for studies to characterize the origin and fate of different samples; this is because they are widespread components of the environmental carbon cycle and are highly resistant to biochemical degradation and diagenesis in the sedimentary record [81].

n-Alkanes is a group of non-polar and photo catalytically stable organic compounds and their bulk characteristics in suspended particle extract can be used to identify two major sources: biogenic and anthropogenic, and provide useful information for the identification of particle sources [76].

In particular, two parameters are mainly relevant as the chemical signature:

- the chain length, the average value and maximum carbon number (C_{max});
- the abundance distribution of the odd/even terms of the series.

One common parameter derived from this predominance is the carbon preference index, CPI: it is computed as the ratio of the sum of odd carbon number n-alkanes vs. the sum of even carbon number n-alkanes [32, 82].

The CPI is a key diagnostic parameter to determine the biogenic and anthropogenic nature of sources of n-alkanes: hydrocarbons composed of a mixture of compounds originating from terrestrial plant material show a predominance of odd-numbered carbon chains with $CPI \approx 5 - 10$ [83, 84] whereas petrogenic inputs have a CPI approximating 1.0 [85–88]. CPI values close to one are also thought to indicate greater input from marine microorganisms and/or recycled organic matter [89].

The CPI has proved of great value in environmental and paleo-environmental biomarker-based research in qualitatively and semi-quantitatively apportioning sources of hydrocarbons found in aquatic sediments: the n-alkane distribution pattern is a biomarker which proves helpful in tracking the origins of organic inputs (biogenic or anthropogenic) and identifying *hot spots* of hydrocarbon contamination [90].

In petrochemistry, n-alkanes are important constituents of petroleum crudes and their transformation products and thus they are useful tools in oil-oil correlation studies because they provide information regarding an oil, its source rock, genetic associations and alteration [91]. In organic geochemistry, the CPI is used to indicate the degree of diagenesis of straight-chain geolipids, and to numerically represent how much of the original biological chain length specificity is preserved in geological samples [85, 86, 90].

Moreover, the chemical characterization of n-alkane constituents of leaf wax coatings has proved to be a quick, reliable and inexpensive method for assessing preliminary chemotaxonomic relationships for systematic classification of plant groups, in combination with other chemical and molecular data: the chemotaxonomic significance of wax alkanes has been demonstrated in studies of many plants groups [92–96].

CPI is strictly correlated to the R index calculated through the $EACVF_{tot}$ computation [32]. The CPI can be calculated by using the different n-alkane terms present in the mixture to describe the different nature of the n-alkane component of the sample [82, 97]. The whole range of n-alkanes is used to describe the whole n-alkane component:

$$CPI_{tot} = \frac{\sum C_{13} - C_{35}}{\sum C_{12} - C_{34}} \tag{3.1}$$

To describe the petrogenic fraction, only C_{12} to C_{25} n-alkanes have to be considered (CPI_{pet}). The heavier $C_{25} - C_{35}$ n-alkanes are used to describe the biogenic contribution (CPI_{bio}).

The R value, which is based on the mean peak height of odd vs. even terms (equation 2.38), can be properly used to estimate CPI: the contribution of selected n-alkanes can be identified by computing the $EACVF_{tot}$ over a partial region of the chromatogram which has been correctly chosen so that it contains a specific range of n-alkanes [32, 98].

3.2.2 Carboxylic Acids

Dicarboxylic acids are among the most abundant organic constituents of ambient particulate matter [75]. Discussion of their potential as tracers for secondary organic aerosol dates back to some of the earliest mass spectral observations of atmospheric aerosols [99].

Dicarboxylic acids are formed in the atmosphere from gas phase photochemical reactions involving a wide range of both anthropogenic and biogenic precursors. They have been identified in smog chamber experiments as atmospheric oxidation products of cyclic olefins [100,101], and proposed as atmospheric oxidation products of aromatic hydrocarbons, fatty acids, and larger dicarboxylic acids [102]. Their aqueous phase formation in cloud and fog water is also plausible [103] and may be linked with photochemically generated radicals [104]. The relatively high concentrations of dicarboxylic acids and their identification as atmospheric reaction products from a variety of different precursors make it useful to investigate their potential as indicators of secondary organic aerosol formation.

Dicarboxylic acids are an important group of water-soluble organic compounds (WSOC) in the atmospheric aerosols [48,75,100,102,105]. They have received much attention because of their potential roles in affecting the global climate. Because of the low vapor pressures and high water solubility, diacids have an influence on the chemical and physical properties of aerosols [106]. Consequently, they may have direct and indirect effects on the earth's radiation balance by scattering incoming solar radiation, which counteracts the global warming caused by the increase of greenhouse gases [107].

Among these dicarboxylic acids, oxalate is generally the most abundant, followed by malonate and succinate in atmospheric aerosols [102,108,109]. Total diacids account for about 1-3% of the total particulate carbon in the urban areas and even above 10% in the remote marine environment [102,109–112].

The use of atmospheric dicarboxylic acids as indicators of secondary formation is complicated by the occurrence of both biogenic and anthropogenic primary sources. Biogenic sources include plant emissions of metabolic products and soil particles. Anthropogenic sources include exhausts from gasoline and diesel powered automobiles [113].

In particular, low molecular weight dicarboxylic acids ($C_3 - C_9$) may yield relevant information on the source strength of anthropogenic vs. biogenic precursors [102,107,113–115]. It has been suggested that the C_3/C_4 ratio is an indicator of enhanced photochemical production of dicarboxylic acids in the atmosphere since succinic acid (C_4) is a precursor of oxalic (C_2) and malonic (C_3) acids. On the other hand, the C_3/C_4 ratio has been used as an indicator of the relative source strength of anthropogenic and biogenic diacid precursors: adipic acid was proposed as a product of the oxidation of anthropogenic cyclohexen, while azelaic acid was thought to come from the oxidation of biogenic unsaturated fatty acids [102, 107, 113, 116].

To date, GC-MS is the method of choice for characterizing individual organic compounds within aerosol samples, primarily because of its high sensitivity and resolving power. The high polarity and low levels (approximately $1ng/m^3$) of these compounds pose special challenges for their identification and quantification because they must first be derivatized, converted to less polar compounds, before they can be eluted through a GC column [102, 113–128].

Two derivatization processes [129] are the ones mainly used to analyze dicarboxylic acids in PM samples because they offer easy sample preparation and display good analytical characteristics:

- esterification of the acid groups using methanol or 1-butanol as derivatizing agent in the presence of a relatively strong acid (BF_3 or BCl_3) (Figure 3.2) [102, 107, 116, 118, 119, 122, 126, 130];
- silylation based on a silylation reagent N,O-bis(trimethylsilyl)-trifluoroacetamide (BSTFA) to form trimethylsilyl (TMS) derivatives (Figure 3.3 [4, 102, 114, 121, 124–126, 128, 131].

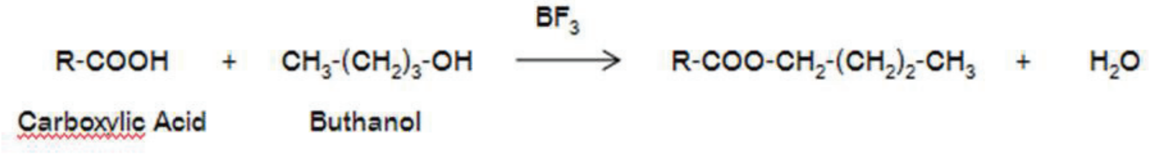


Figure 3.2: BF_3/Buthanol Derivatization Procedure

The two methodologies differ in terms of the stability of the derivatives formed, the presence of interfering by-products and speed. Moreover, a combination of the two procedures has been employed to yield a multistep derivatization by which -COOH groups are initially derivatized with BF_3/alcohol and then the remaining hydroxy or keto groups are silylated with a sylilation

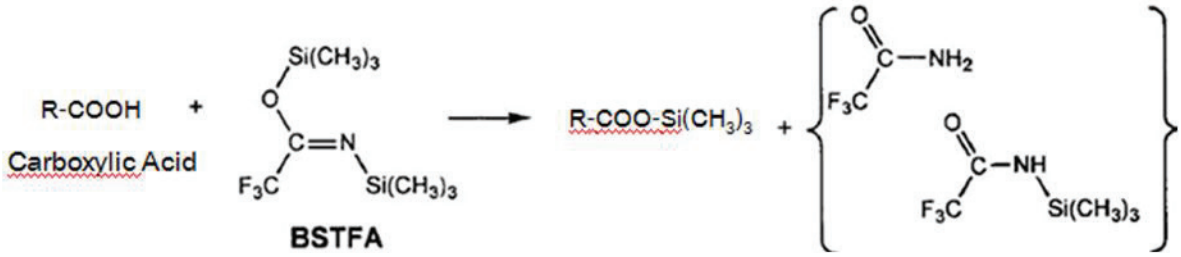


Figure 3.3: BSTFA Derivatization Procedure

reagent [126].

Not only the dicarboxylic acids are markers for PM studies, fatty acids (long-chain monocarboxylic acids) have been studied too. Fatty acids are emitted to the atmosphere from many sources: the lower molecular weight n-alkanoic acids ($< C_{20}$) are mainly emitted by petroleum based sources and meat cooking, while the heavier $C_{20} - C_{30}$ terms, which display a strong even-to-odd carbon number preference, are mostly derived from plant waxes [116].

C_{16} (hexadecanoic) and C_{18} (octadecanoic) saturated acids are the two most abundant in the PM [102, 116], and they accounted for 50-70% of the total fatty acids. The strong even carbon number predominance ($CPI > 10$) suggests that the fatty acids are mainly biogenic [132].

4

Results and Discussion

Contents

4.1 N-alkanes 34
4.1.1 Application to samples of known origin 34
4.1.2 Application to real samples of PM 39
4.2 Carboxylic acids 43
4.2.1 Low Molecular Weight (LMW) dicarboxylic acids 43
4.2.2 N-alkanoic acids 48

The developed method's robustness and reliability in estimating the n_{max} and R parameters were verified on simulated chromatograms with a known distribution of the sequence terms. All the results obtained from the $EACVF_{tot}$ calculation and from the $EACF$ plot inspection, show a good agreement between the theoretical values and the calculated ones [32].

The attention of the present thesis has been mainly focused on chemical characterization of environmental complex samples, in particular the applicability of the $EACVF$ method was tested on experimental chromatograms of samples of known origin (anthropogenic or biogenic), such as oil samples and plant extracts. The parameters obtained are useful molecular markers for comparing known sources and observed atmospheric samples to identify sources of organic matter emissions [32].

4.1 N-alkanes

4.1.1 Application to samples of known origin

As a preliminary step of this study, the reliability of the method has been tested on real samples of known origin.

Gasoil Sample (Anthropogenic origin) - As an example, the GC-MS signal of the volatile components of a commercial diesel fuel was studied (Figure 4.1a): the SIM signal for monitoring the n-alkanes at m/z values of 57, 71 and 85 is reported. The n-alkanes were identified using the GC retention times of the reference standards ($C_{10} - C_{30}$): the main components are mid-chain n-alkanes $C_{10} - C_{25}$, C_{17} and C_{19} being dominant.

A visual examination of the chromatogram shows a typical chromatographic profile of petrogenic n-alkanes characterized by no odd-to-even predominance. The $EACVF_{tot}$ was computed on the whole chromatogram (lower solid line in Figure 4.1c): its plot clearly shows a monomodal distribution of the $EACVF_{tot}$ peak height suggesting a homogeneous distribution of the odd/even terms. Such a pattern can be quantified by computing CPI_{tot} according to equation 2.43: by selecting the proper retention region containing $C_{13} - C_{25}$ n-alkane ranges, CPI_{pet} (petrogenic) can be estimated to characterize the petrogenic fraction present in the sample: CPI_{tot} and CPI_{pet} values close to 1 were obtained (estimated values, 2^{nd} and 3^{rd} columns in Table 4.1). With the developed algorithm the n_{max} n-alkanes present in the sample can be directly estimated from the $EACVF_{tot}$ peaks at $\Delta t = bk$, even k [32] (estimated values, 5^{th} column in Table 4.1).

The accuracy of the results was checked by comparing them with results obtained using the traditional procedure. It requires identification of the n-alkanes by comparison to reference standards and MS spectra, integration of the identified peaks, computation of CPI as a ratio of the sum of concentrations of the odd-numbered carbon alkanes vs. that of the even-numbered terms. The obtained results (traditional calculations, 6^{th} and 7^{th} columns in Table 4.1) show a close similarity with data estimated by $EACVF_{tot}$: this agreement is a proof of the usefulness of the procedure for a simple and quick characterization of the n-alkane distribution pattern as a molecular biomarker in complex samples.

The ability of the $EACVF_{tot}$ procedure to handle complex signals can be emphasized by extending the investigation to involved TIC signals. The TIC chromatogram of the oil sample was studied (Figure 4.1b): it displays the typical chromatographic profile characterized by the UCM band (Unresolved Component Mixture) formed by a cluster of unresolved peaks. The $EACVF_{tot}$ plot (Figure 4.1c, upper bold line) is strongly affected by the specific pattern of the UCM band which is superimposed on the deterministic $EACVF_{tot}$ peaks, displaying monomodal height distribution.

Nevertheless, the $EACVF_{tot}$ model makes it possible to single out the n-alkane sequence properties by computing the n_{max} and CPI_{tot} values using equations 2.45 and 2.43 on $EACVF_{tot}$ computed over the whole original signal. The obtained results (4^{th} row in Table 4.1) show a close similarity to the data obtained from the SIM signal (3^{rd} row in Table 4.1) and from the traditional calculation method. This result confirms the robustness of the developed method in extracting reliable information from the direct handling of complex chromatograms, such as SIM and TIC GC-MS involved signals [32].

Plant samples (Biogenic origin) - Dichloromethane extract of flowers of *Mimosa* plant was submitted to GC-MS analysis: the SIM chromatogram was monitored at $m/z = 57 + 71 + 85$ to represent the aliphatic hydrocarbon fraction. The chromatogram of hydrocarbons extracted from *Mimosa* flower are pictured in Figure 4.2a. The main components are mid- and long-chain n-alkanes $C_{21} - C_{33}$. C_{23}, C_{25}, C_{27} and C_{29} are the dominant long-chain n-alkanes in the GC profiles. A visual examination of the chromatogram shows a typical chromatographic profile of n-alkanes from vascular land plants characterized by a high odd-to-even predominance of long chain $C_{25} - C_{35}$ with $CPI \approx 5 - 10$.

The $EACVF_{tot}$ plot computed on the whole chromatogram (Figure 4.2b) clearly shows a bimodal distribution of the $EACVF_{tot}(bk)$ peak height with lower values at odd k values (combination term, equation 2.37) and higher values at even k (addition term, equation 2.36). Such a pattern is diagnostic of an odd/even prevalence that can be quantified by computing CPI_{tot} according to equation 2.43: a CPI_{tot} value close to 5 was estimated (estimated values, 5^{th} row in Table 4.1). To better characterize the plant chemical composition, the CPI_{bio} index was also computed by selecting the chromatographic region containing the long chain $C_{24} - C_{33}$ terms (estimated values, 4^{th} column in Table 4.1). The developed algorithm also yields an estimation of the n_{max} n-alkanes present in the sample (estimated values, 5^{th} column in Table 4.1) directly from $EACVF_{tot}$ peaks at $\Delta t = bk$ even k (equation 2.45).

To check the accuracy of the obtained results, the traditional procedure based on calculation on each identified and integrated peak was applied to compute CPI_{tot}, CPI_{bio} and n_{max} values (traditional calculations, $6^{th} - 9^{th}$ columns in Table 4.1). The close similarity between the computed and the estimated $EACVF_{tot}$ data proves the reliability of the developed method to identify and characterize the abundance distribution of biogenic n-alkanes, and this may also be useful in extracting information for a chemotaxonomic approach [32].

	$EACVF_{tot}$ Estimation				Traditional Calculation			
	CPI_{tot}	CPI_{pet}	CPI_{bio}	n_{max}	CPI_{tot}	CPI_{pet}	CPI_{bio}	n_{max}
Fuel (SIM)	1.60	1.67	-	19.4	0.97	0.96	-	20
Fuel (TIC)	1.52	1.56	-	19.0	1.03	1.00	-	20
Mimosa	5.30	-	5.80	12.4	5.80	-	4.67	11

Table 4.1: Parameters of real samples [32]

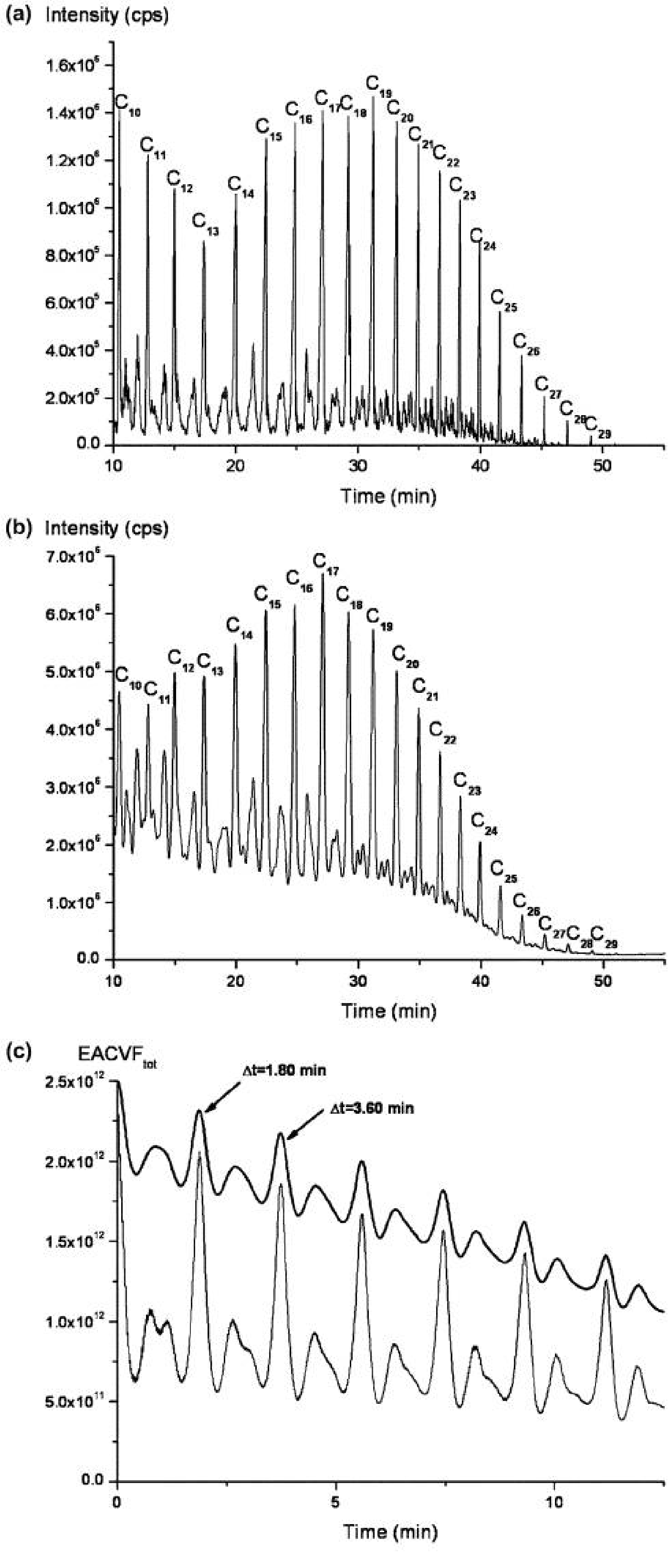


Figure 4.1: GC-MS chromatograms (SIM and TIC) and relative $EACVF_{tot}$ plots of a Fuel sample [32]

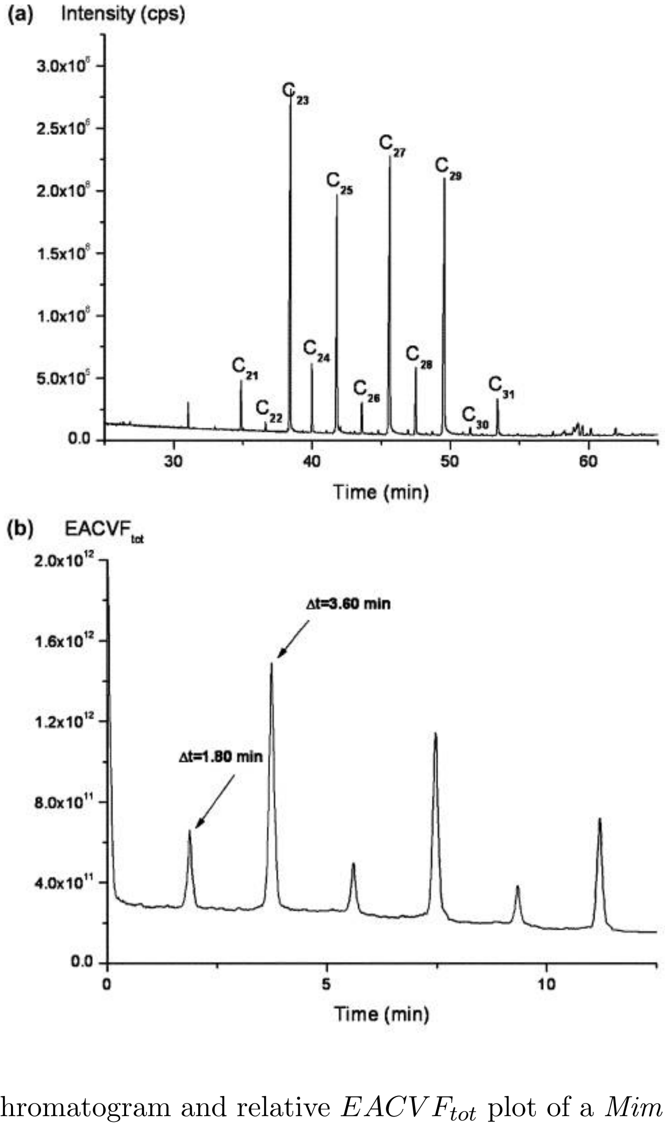


Figure 4.2: GC-MS chromatogram and relative $EACVF_{tot}$ plot of a *Mimosa* flower extract [32]

4.1.2 Application to real samples of PM

The developed signal processing procedure, based on the AutoCoVariance Function computation, was applied to handle GC-MS signals of PM samples collected in northern Italy (Milan): thanks to the method, information on the relative contribution of the homologous series, $series\%$ (percentage of homologous series terms over the total chromatogram area) and the CPI values, can be directly estimated from the $EACVF_{tot}$ and the PC computation takes just a few minutes [98].

The chemometric method was applied to all chromatographic signals from GC-MS analysis of the PM samples. The aim was to test the method's ability to characterize the n-alkane contribution, in terms of CPI, and its applicability as an high-throughput method for analysis of the huge amounts of data from environmental monitoring.

After the GC-MS analysis, the n-alkanes series terms, ranging from C_{14} to C_{32} can be identified in the investigated PM samples. However, the lighter $C_{14} - C_{19}$ terms were found at a low concentration level, lower than the detection limit for more than 50% of the samples. The first region of the chromatogram, where $C_{14} - C_{19}$ n-alkanes elute, was quite disturbed due to coelution of other interfering compounds. Moreover, the lighter n-alkanes with $C \leq 19$ are generally considered too volatile to be accurately determined in PM samples as they incur evaporative losses during the sampling and analytical procedures [133–135].

For all the above reasons, the terms ranging from C_{20} to C_{32} were investigated as potential tracers for biogenic/antropogenic emissions: they were detected in all the analyzed PM samples, displaying a concentration level higher than the detection limit for most of the samples ($> 80\%$).

The first step of data handling consisted of a procedure to linearize the chromatographic signal to obtain constant retention increments between subsequent terms of the homologous series (see section 2.3). The Autocovariance Function was then numerically calculated from the linearizated chromatogram, according to equation 2.1. Then, the MATLAB® algorithm has been used to directly estimate the parameters n_{max} and CPI_{EACVF} from the EACVF computed on a properly selected region of the chromatogram corresponding to the $C_{20} - C_{32}$ n-alkanes [98].

The EACVF method was tested and compared with the traditional one for 22 samples (see Table 4.2) The CPI_{Trad} parameter was computed using the traditional procedure based on peak integration of the $C_{20} - C_{32}$ n-alkane GC-MS signal to describe their abundance distribution [82]

(2^{nd} column in Table 4.2). Most of the analyzed samples show CPI ≈ 1 values indicating strong contribution of emissions from urban *winter* sources, such as domestic heating (for example natural gas, oil, and wood combustion) generating a random distribution of odd/even terms of the series. On the other hand, the summer samples show higher CPI values (1.5-3.5) due to the higher contribution of the odd terms C_{27}, C_{29} and C_{31} originating from plant material which yield maximum emissions during the vegetative season [136, 137].

The EACVF was directly computed on the GC-MS signal (SIM signal at m/z values of 57, 71 and 85): the region 30-60 min was selected, since it contains the $C_{20} - C_{32}$ n-alkanes (Figure 4.3, sample **MI-17**). In comparison with the complex original GC signal, the EACVF plot (Figure 4.4) shows a simplified pattern characterized by a sequence of deterministic peaks located at $\Delta t = 2.8min$, the retention incrementen between subsequent terms of the the n-alkane series ($\Delta t = b$, equation 2.10) under the experimental GC conditions used. Such EACVF peak is diagniostic, directly identifying the presence of n-alkanes and hence there is no need to compare them with the GC retention times of the reference standards ($C_{20} - C_{32}$). The main information on n-alkanes series are directly extracted from the values of the EACVF computed at $\Delta t = bk$, for chracteristic k values. The number n_{max} of n-alkanes present in the sample can be directly estimated from the EACVF peaks at $\Delta t = bk$ for even k: for all the investigated GC-MS signals the n_{max} values were correctely estimated as $n_{max} = 13$. The abundance distribution of the odd/even terms, can be quantified by computing CPI_{EACVF} values directly from EACVF using

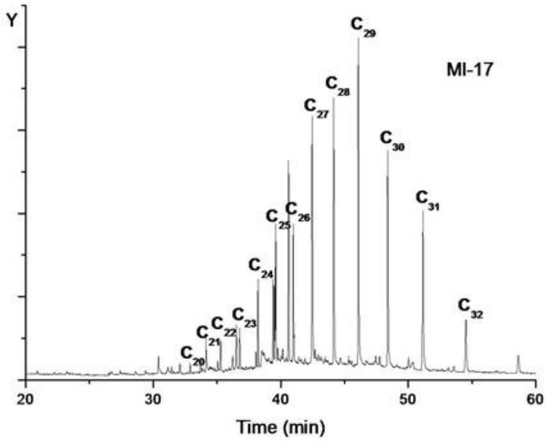


Figure 4.3: GC-MS chromatogram of the Sample **MI-17** [98]

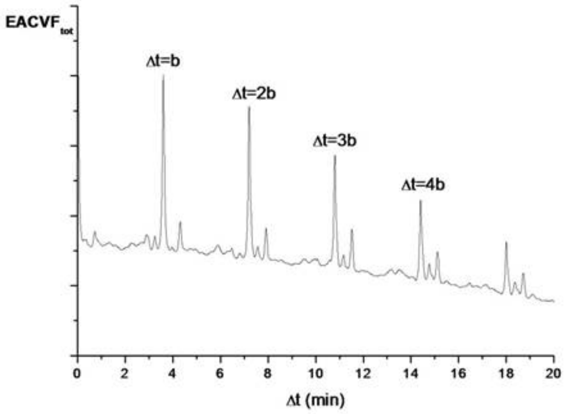


Figure 4.4: $EACVF_{tot}$ plot of the Sample **MI-17** [98]

the values at $\Delta t = b$ and $\Delta t = 2b$ (CPI_{EACVF} , 3^{rd} column in Table 4.2).

The accuracy of the obtained results was checked by comparing the CPI_{EACVF} with CPI_{Trad} values obtained using the two procedures (CPI_{Trad} vs. CPI_{EACVF}, 2^{nd}, 3^{rd} columns in Table 4.2) and estimating the relative estimation error ($\epsilon\%$, 4^{th} column in Table 4.2).

In general, a good agreement was achieved between the two procedures: the relative error $\epsilon\%$ was lower than 15% for 70 of the 76 investigated samples and lower than 5% for 22 of them. The exceptions are 6 PM samples for which different CPI_{Trad} and CPI_{EACVF} values were estimated ($\% \geq 15\%$). They correspond to the samples containing the lowest n-alkane abundance and which generate complex GC signals showing coeluting components and superimposed UCM band: this makes it very difficult to integrate the n-alkane peaks and to deconvolute the EACVF plot to prevent an unbiased estimation of the CPI_{Trad} and CPI_{EACVF} values.

Many investigated signals display a high contribution of the UCM hump because, in an effort to obtain a fast analytical procedure for n-alkane determination, the samples were obtained by a simple solvent extraction, without any extract purification. These conditions may yield an ambigous n-alkane characterization as a consequence of the coelution yielding of complex superimposed signals: this is particularly true for samples containing low abundance n-alkanes. These chromatograms were handled with the complete procedure for deconvolving the UCM contribution from $EACVF_{tot}$ (* data in the Table 4.2). It must be noted that the whole procedure is reliable in estimating accurate parameters, since it also includes retention time rescaling and UCM component subtraction.

Sample	CPI_{Trad}	CPI_{EACVF}	$\epsilon\%$
MI-1	1.48	1.70*	14.50
MI-2	1.47	1.02*	30.50
MI-3	1.27	1.36	7.10
MI-4	1.38	1.00*	27.60
MI-5	1.61	1.87*	15.90
MI-6	1.43	1.60*	11.70
MI-7	1.56	1.50*	3.95
MI-8	1.48	1.38	6.67
MI-9	1.20	1.05	12.50
MI-10	1.64	1.48	9.75
MI-11	1.69	1.45*	14.20
MI-12	1.89	1.71*	9.60
MI-13	1.88	1.61*	14.10
MI-14	1.31	1.17*	10.70
MI-15	1.77	1.71*	3.25
MI-16	1.59	1.62	1.64
MI-17	1.23	1.13	8.10
MI-18	1.76	1.53*	13.00
MI-19	1.68	1.54*	8.20
MI-20	3.01	2.70*	10.20
MI-21	3.38	2.98*	11.90
MI-22	1.18	1.14	3.31

Table 4.2: Comparison between CPI values calculated through the traditional method and the $EACVF_{tot}$ one [98]

4.2 Carboxylic acids

Carboxylic acids in general, represent one of the biggest part of the organic fraction of particulate matter. There are different types of acids; low molecular weight acids (LMW) are the water soluble part of the carboxylic acids and they are usually studied in source apportionment investigation. In a GC-MS analysis they are hard detectable compounds, because of their requirement of preliminary derivatization step.

N-alkanoic acids are very useful to characterize and to identify the biogenic contribution to environmental pollution, through the investigation of the homologous series (usually the $C_{14} - C_{24}$ terms) and the calculation of the relative CPI values.

4.2.1 Low Molecular Weight (LMW) dicarboxylic acids

The determination of LMW dicarboxylic acids ($C_3 - C_9$) is very important because they contain relevant chemical information to distinguish primary vs. secondary sources as well as anthropogenic vs. biogenic precursors. Preference was given to a faster one-step derivatization procedure to determine selected target compounds: the advantages and drawbacks of the methods using BF_3/alcohol and $BSTFA$ are investigated and compared in terms of precision and accuracy of the results, sensitivity and detection limit of the procedure [129].

BF_3 esterification - The BF_3/alcohol reagent converts either carboxyl groups into butyl esters or aldehyde groups into dibutyl acetals [138, 139]. Starting from the original Kawamura paper [130], different modifications have been reported and widely applied to make BF_3/alcohol derivatization the most widely used procedure for determining LMW oxygenates in atmospheric samples [102, 107, 116, 118, 119, 122, 126]. In particular the BF_3/butanol procedure has distinct advantages for quantifying LMW compounds because the resulting butyl derivatives are less volatile and more resistant to evaporative losses than the BF_3/methanol scheme [118, 119]. Because of the presence of residual acid, the products cannot be directly injected into the GC, rather a purification step is required before injection [113]. The distinct advantage is that environmentally safe esters are formed.

Silylation - Silylation is another common derivatization technique used to derivatize polar compounds prior to GC-MS analysis. The usual reagents for PM analysis are trimethylchlorosilane (TMCS), N-methyl-trimethylsilyltrifluoroacetamide (MSTFA), N,O-bis-(trimethylsilyl)trifluoro acetamide (BSTFA) and N-(t-butyldimethylsilyl)-Nmethyltrifluoroacetamide (MTBSTFA) [4, 102, 114, 121, 124–126, 128, 131]. During the silylation reaction, all the hydroxyl groups are converted into their corresponding trimethylsilyl derivatives via a substitution reaction which yields one main product for each compound and with high conversion efficiency [4, 128].

The reaction is low moisture sensitive and requires mild conditions to complete the derivatization needed to achieve GC-MS detection at very low concentrations [4, 128, 131]. In opposition to alkylation, silylation normally does not require a purification step and the derivatives can be injected directly into the GC system [4, 113, 128, 131]. However, it presents some drawbacks, such as the fact that the silylation reagent is dangerous and some artifacts can be produced in the reaction [126].

The two most common derivatization procedures were compared for quantitative analysis of dicarboxylic acids in PM samples by focusing attention on two challenging conditions:

1. Quantification of lighter C_3 and C_4 dicarboxylic acids, since they contain relevant information for source apportionment and secondary organic aerosol formation [102, 107, 113–116, 118];
2. analysis of PM samples collected by low-devices ($55m^3$ air volume sampled over $24h$) requiring the highest method sensitivity at the trace level.

The method sensitivity and linearity were evaluated by computing calibration curves with standard solutions [129]. Different experimental derivatization conditions have been widely applied to derivatize LMW oxygenate compounds for subsequent GC determination in PM samples [114, 115, 121, 124–126, 128]. An optimization study was performed on the derivatization conditions that most affect analytical responses: reaction temperature and duration time. This study brought to the following reaction conditions: 75°C (reaction temperature) and 90 min (reaction time) [129]. The precision and accuracy of the procedure were assessed through three replicate measurements of blank quartz fiber filters spiked with $C_3 - C_9$ target acids at three concentration levels. Good recoveries were found for all the target compounds ranging from 78% for malonic acid to 115% for azelaic acid. The procedure also displays good reproducibility as evaluated by RSD% values on three replicates lower than 10% [129].

The obtained results confirm that both the methods are rapid, reproducible, trace level procedures suitable for environmental monitoring of dicarboxylic acids. However, some differences can be singled out when the two procedures are compared for the challenging application of quantitative determination of lighter $C_3 - C_9$ di-carboxylic acids at trace levels.

In general, the analytical response obtained for BSTFA derivatives was higher than the one for the butyl esters. This can be explained by the silylation reaction yield or by the stability of the derivatives during handling. As a consequence, the silylation procedure displays higher sensitivity with lower detection limit values for all the investigated $C_3 - C_9$ di-carboxylic acids, compared to butyl esterification. On the other hand, the sensitivity of the BF_3/BuOH method strongly depends on the acid molecular weight: it is unreliable for the lower $C_3 - C_4$ terms and it significantly increases with the acid molecular weight to achieve detection limits comparable to those of silylation for the heavier $C_7 - C_9$ acids. The low sensitivity for $C_3 - C_4$ acids is also due to the concomitant higher volatility of their derivatives which yields evaporative loss during the derivatization procedure.

In order to confirm the findings obtained upon standard solutions, the two methods were applied to environmental PM matrices. To obtain comparable results on the same aerosol sample, each sample (PM sample 1, 2 in Table 4.3) was obtained by two consecutive sets of samples (2 quartz fiber filters) combined for extraction and then halved to separately perform derivatization prior to GC-MS analysis. The TIC chromatogram of the BF_3/BuOH derivatized sample (Sample 1) is reported in Figure 4.5: the SIM signals at specific m/z values for the $[M - 73]^+$ ions of each butyl ester were selected for identification and quantification of the target dicarboxylic acids. The silyl derivatives obtained with BSTFA reagents on the same aerosol sample (Sample 1) were analyzed under SIM detection mode at $m/z = 74 + 147 + 149$ (SIM chromatogram in Figure 4.6).

The concentrations of the target dicarboxylic acids were measured with both the procedures using the calibration curves: the obtained results are reported in Table 4.3 for both the samples.

As verified on standards, the lighter C_3 and C_3 acids escaped detection by the BF_3/BuOH method, because of their high susceptibility to evaporative loss. For the other acids, a good agreement (within 4%) was found between the results obtained from the two procedures: this result proves that both derivatization procedures produce repeatable quantification of the target acids for the investigated samples, even when operating at, or close to, their detection limits.

Similar abundance was found for the individual species, independently of the carbon chain length, with malonic and azelaic acids predominant. These results are consistent with literature

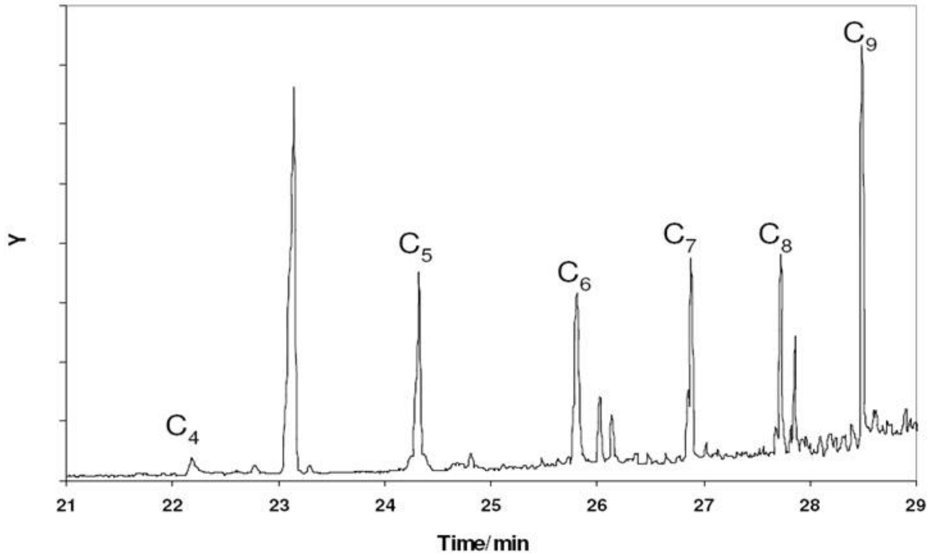


Figure 4.5: GC-MS chromatogram of a PM sample (BF_3/BuOH derivatization procedure) [129]

on dicarboxylic acids in $PM_{2,5}$ for a rural sampling site [117, 121, 124, 125, 128]. The predominance of the C_9 diacid is expected since it is an oxidation product of biogenic unsaturated fatty acids. Accordingly, a low value, close to 0.5, was computed for the C_6/C_9 ratio to indicate a high biogenic input for aerosol diacids: 0.53 and 0.54 for both samples and using both procedures.
Moreover, BSTFA derivatization also makes it possible to compute the C_3/C_4 ratio as another marker of diacid origin: both samples yield a value of 1.3 as it is commonly observed in atmospheric aerosols with low anthropogenic sources (combustion of fossil fuels produces $C_3/C_4 \approx 0.35$) and reduced photo-induced secondary formation of dicarboxylic acids (that would yield higher $C_3/C_4 \geq$ values) [102, 107, 113, 115, 116].
The BSTFA procedure is preferable when comparison is performed under the most limiting conditions concerning analysis of lighter $C_3 - C_4$ terms in PM filters collected by low volume air samplers: it provides lower detection limits ($\leq 3ng/m^3$) and higher reproducibility (RSD% $\leq$ 10%).

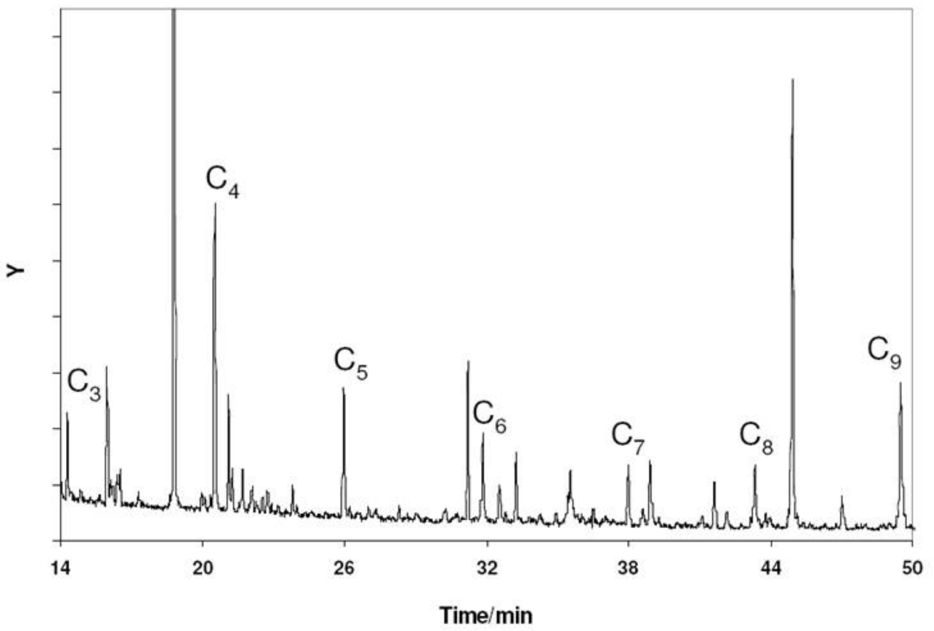


Figure 4.6: GC-MS chromatogram of a PM sample (BSTFA derivatization procedure) [129]

	BF_3/BuOH derivatization		BSTFA derivatization	
	Sample 1	Sample 2	Sample 1	Sample 2
Malonic Acid (C_3)	–	–	5.2 ± 1.4	5.4 ± 1.5
Succinic Acid (C_4)	–	–	3.8 ± 1.2	4.0 ± 1.6
Glutaric Acid (C_5)	2.6 ± 3.3	2.8 ± 3.2	2.5 ± 1.6	2.7 ± 1.4
Adipic Acid (C_6)	3.1 ± 2.0	3.3 ± 3.8	3.0 ± 1.2	3.2 ± 1.6
Pimelic Acid (C_7)	2.7 ± 1.8	3.0 ± 2.6	2.6 ± 1.8	2.9 ± 1.4
Suberic Acid (C_8)	2.6 ± 2.7	3.0 ± 3.9	2.5 ± 1.4	2.9 ± 1.3
Azelaic Acid (C_9)	5.8 ± 2.8	6.1 ± 3.4	5.6 ± 1.8	6.0 ± 2.0

Table 4.3: Concentrations (reported in ng/m^3) of the target dicarboxylic acids measured on two experimental PM samples after derivatization [129]

4.2.2 N-alkanoic acids

The $EACVF_{tot}$ method was also applied to characterize n-alkanoic acids, as an homologous series of organic components useful in discriminating the relative extent to which various sources contribute to the aerosol burden of organics. After derivatization based on a BSTFA procedure [4, 129], the urban and rural PM samples were submitted to GC-MS analysis: the n-alkanoic acids present in the samples were identified in the SIM signal monitoring the typical fragments of the TMS derivatives at $m/z = 75 + 147$ (Figure 4.7).

Under the experimental conditions used [132], the retention increment for subsequent n-alkanoic acids is $b = 2.5min$. The $EACVF_{tot}$ was computed on the whole signal (Figure 4.8: solid line): deterministic peaks at $\Delta t = 2,5min$ and multiple values are diagnostic for the presence of this homologous series. All the data set to characterize the series are estimated (Table 4.4, $2^{nd} - 5^{th}$ columns, $EACVF_{tot}$ estimation) and compared to results obtained with the traditional procedure (Table 4.4, $6^{th} - 9^{th}$ columns, traditional calculations). The $EACVF_{tot}$ plot shows a marked bi-modal distribution with a predominant peak at $\Delta t = 2,5min = 2b$: this is consistent with predominant contribution of hexadecanoic and octadecanoic acids that are known to be the most abundant species in most of the PM samples [102, 116]. The even/odd prevalence of

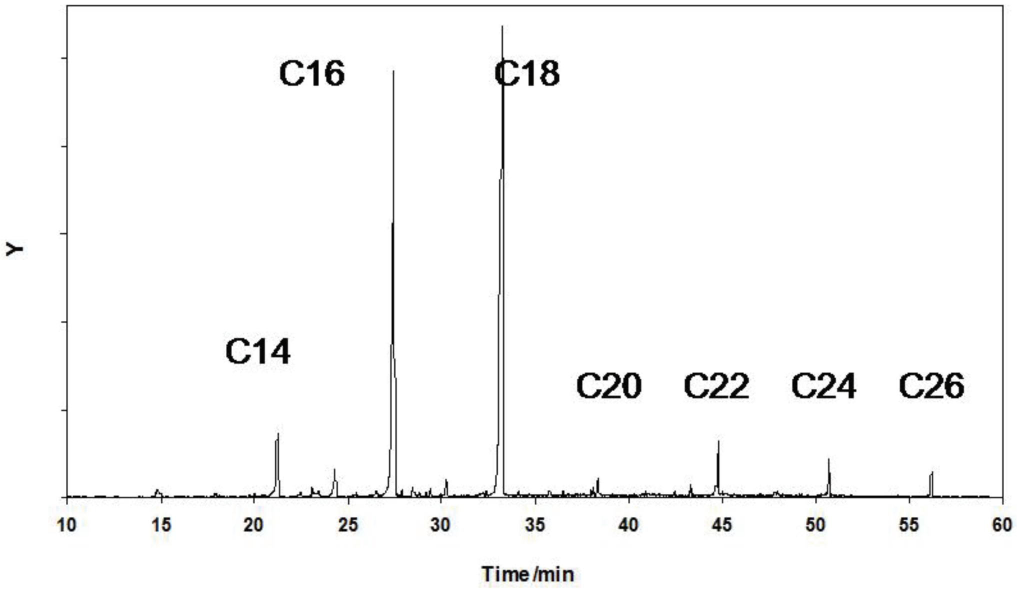


Figure 4.7: GC-MS chromatogram (n-alkanoic acids) of $PM_{2.5}$ rural sample [132]

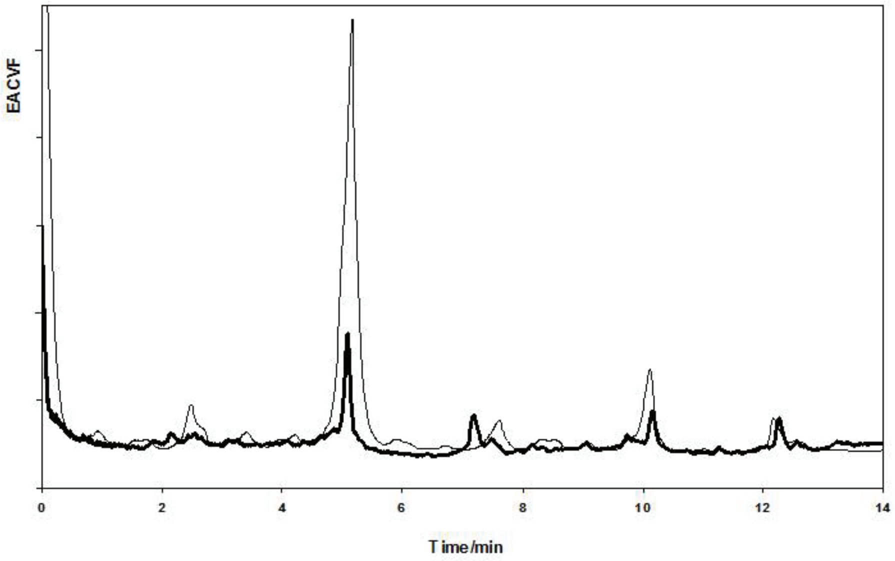


Figure 4.8: $EACVF_{tot}$ plot of n-alkanoic acids homologous series in a $PM_{2.5}$ rural sample [132]

acid isomers was confirmed by high $CPI = 9.8$ value (Table 4.4).

To extract information on the biological sources of n-alkanoic acids, the selected chromatographic region containing the $C_{20} - C_{26}$ terms was separately investigated by computing $EACVF_{plant}$. The obtained $EACVF_{plant}$ plot (Figure 4.8, bold line) clearly identifies the contribution of biogenic sources, since it displays the strong bi-modal distribution ($EACVF_{plant}(bk)$ peaks are low for odd k and high for even k) characteristic of a strong odd/even prevalence. This is confirmed by the high CPI value ($CPI_{plant} = 18.7$) computed from subsequent $EACVF_{plant}$ peaks, reflecting the stronger vascular plant wax signatures.

The contribution of biogenic n-alkanoic acids in PM samples can also be directly estimated by the ratio between $EACVF_{tot}(5min)$ and $EACVF_{plant}(5min)$ computed on each chromatogram: the plant fraction ($\geq C_{20}$ congeners) accounted for about 25% of the total measured n-alkanoic acids levels in the rural sample [132].

	$EACVF_{tot}$ Estimation				Traditional Calculation			
	n_{max}	CPI_{tot}	n_{plant}	CPI_{plant}	n_{max}	CPI_{tot}	n_{plant}	CPI_{plant}
$PM_{2,5}$ **(rural)**	13.6	9.8	7.5	18.7	14	9.4	8	17.2

Table 4.4: Parameters of a rural $PM_{2,5}$ sample (n-alkanoic acids series) [132]

5

Conclusions

The EACVF procedure developed in this Ph.D. project proves to be a simple data processing method to efficiently handle a multicomponent chromatogram in order to characterize the chemical composition of the complex sample. The method is particularly powerful in identifying the presence of an ordered sequence of compounds, singling it out from the complexity of the disordered chromatogram; the two components, ordered and disordered, can be separated and quantitatively evaluated; that is, number of compounds of each pattern can be estimated. Such information can be extracted by handling the simple GC signal, without any information on the chemical structure of the components.

Moreover, the power of the method is significantly magnified if combined with the SIM detection. In fact, the EACVF essentially singles out structure retention correlation of thermodynamics origin, whereas SIM provides further selectivity to the method related to selected molecular structures.

The information obtained by the method makes it possible to analytically characterize the sample not only in terms of identification and quantification of selected SCs but also of identification and quantification of the specific SC homologous series building up the total mixture. Therefore, the present procedure seems to be not merely a powerful chemometric tool for handling complex chromatograms but also a new approach for a comprehensive characterization of a complex multicomponent chromatogram.

In fact, the EACVF plot can be really considered as a *SC class chromatogram* for the separation, the identification, and the quantification by classes, which is additional information, compared to the overall and sometime indistinct sequence of overlapping peaks.

In comparison with the traditional procedure based on computation performed on integrated chormatographic peaks, the EACVF method displays three fundamental advantages:

- it saves time and labor in data handling, thus increasing throughput and flexibility;
- it increases result reliability by deconvolving complex signals into its components;
- it reduces the subjectivity of human intervention, thus improving data quality [98].

The introduction of the data pre-processing step to linearize retention time axis increases the EACVF method applicability and robustness to investigate complex chromatograms obtained under usual experimental conditions.

The reported results [31] constitute only an example of the wealth of information that can be obtained using the present approach.

The method can be proposed to investigate the chemical composition of complex samples of environmental interest: structural class information can be quite useful in monitoring steps in industrial processes and controlling environmental quality, analyzing biomarkers and environment pollutants in air, water and soil.

The procedure then has been focused on a description of the chemical pattern of n-alkanes homologous series, in particular, on the reliable computation of the CPI_{tot} index, as a descriptor of characteristic n-alkane distributions to be used as a signature of specific organic sources. It has proved suitable for the study of long chain n-alkane distributions dominated by odd carbon-numbered homologs, reliable indicators of terrigenous inputs in environmental and paleo-environmental studies.

In addition, the method may be applicable for diagnostic fingerprinting ratios in relation to forensic oil spill identification or bioindicating of the general degree of environmental pollution.

$EACVF$ procedure has been studied not only for n-alkanes homologous series, but even for n-alkanoic acids. Both the series are present in environmental samples (even if with different percentage values [75]) and they both could be investigated to perform a source apportionment study.

6

Future Perspectives

At present, the application of the developed procedure is limited by fulfilling the severe conditions of the SC concentration, in particular, the average abundance of the SCs belonging to the homologous series compared to that of the total mixture. However, the good data obtained from the benzin sample and the possibility to check their availability by handling the SIM signal seem to be a very promising result concerning the applicability of this simple method to unknown real samples [32]. It is clear that the present procedure must be extended to the general, most usual condition, where the concentration of the homologous series is different from that of the majority of the other SCs. For this aim, further theoretical development and application to real cases are under study.

Moreover, the method robustness toward experimental limitations is under study: how the procedure is powerful in overcoming problems related to experimental chromatograms obtained in unfavorable conditions acquisition, such as nonlinear temperature programming conditions and noisy signals. Another limit of the present procedure may be the high concentration of the sample components yielding overloading effects with consequent peak shape distortions and a wide concentration range (over several orders of magnitude) of the detectable components: EACVF may be mostly affected by the predominant components obscuring the least abundant compounds.

At present, the procedure has been tested on a limited number of classes of compounds, but it is obvious that it is general and can be extended to different classes of compounds, if specific fragments for SIM detection are selected, and to more complex mixtures.

The whole signal processing procedure makes it possible to achieve a systematic characterization of complex samples by compound class (homologous series, polarity, or functionality): this is the only reliable information that can be drawn from many real-world samples, natural, industrial or environmental samples where, because of matrix complexity, the separation of all the components is far from being achieved in 1D separations. In such complex samples, given its suitability for group separations, the best technique for analysis of organics is a two-dimensional system, like GCxGC: it is ideal for complex samples containing thousands of compounds but a relatively low number of chemical classes.

The method may be extended to different homologous series in order to characterize specific organic markers for identifying sources and tracing inputs in the environment. The results directly obtained by computing $EACVF_{tot}$ on the whole multicomponent chromatographic signal can constitute the basis for further data analysis using multivariate statistical methods, such as discriminant analysis (DA), cluster analysis (CA) or principal component analysis (PCA) to gain a better understanding the organic component contribution of inputs in the environment [32].

The method has been further extended to handle the huge amount of data obtained from 2D separations [26, 140, 141]: study of the 2D-EACVF may form the basis for a comprehensive interpretation of the data matrix acquired in full scan GC-MS analysis containing the whole MS information on component chemical structure.

In this case, the bidimensional investigation gives two different kind of information, starting from the 2D-EACVF plot: as for the 1D-EACVF, it's possible to obtain the main parameters of the homologous series, n_{max} and CPI, directly from the Δt profile; but when a GC-MS matrix data is under investigation, another information is obtainable, directly from the $\Delta m/z$ profile: a *fingerprint* of the homologous series present in the sample. In other words, peaks on the $\Delta m/z$ profile give a qualitative information about the series because organic compounds series, such as n-alkanes or n-alkanoic acids, have characteristic $\Delta m/z$ values (for example $\Delta m/z = 14, 28, 42, ...$ for n-alkanes), depending on the characteristic fragments of the mass spectra.

The sample preparation is another critical aspect that could influence the goodness of the analysis results. Usually, a simple solvent extraction, starting from the PM filter, has been used for the sample preparation and, if necessary, a derivatization procedure, directly on the extract. New methods are under study, such as the DTD method (Direct Thermal Desorption). This innovative technique uses a little piece of the filter, directly "injected" in the GC-MS system. A thermal desorption procedure is then executed on the filter and all the components are immediatly injected in the GC column. This DTD method increases the sensibility of the analysis and severely reduces the time of sample preparation.

Appendix

Appendix A

```
function [dref,MZS,YS] = mattia(MZ,Y,spikes,spikesWidth,badSpikes,badSpikesWidth,badMZ)
% INPUT
%  MZ   = array delle ascisse (caso 1D: tempi di ritenzione,
%         caso 2D: spettri di massa)
%         Must be given as column vector or column-major matrix
%  Y    = array delle ordinate (intensità)
%         Assumed to be a nY x nMS matrix
%  spikes = array delle ascisse in cui sono individuati i picchi di
%           intensità da equispaziare
%  spikesWidth = halfwidth of each spike in number of ascissa points.
%        Can be a constant or a vector
%        of the same length as spikes
%  badSpikes (optional) = spikes to be replaced
%  badSpikesWidth = halfwidth of each bad spike in number of ascissa points.
%        Can be a constant or a vector of the same length as badSpikes
%  badMZ (optional) = mass spectra to be replaced
%
% OUTPUT
%  MZS  = array delle nuove ascisse
%  YS   = array delle nuove ordinate con picchi equidistanti

if (size(MZ,1) ~= size(Y,1))
    error('FATAL ERROR: mismatching sizes of MZ and Y');
end
[nY,nMS] = size(Y);
% eventuale rimozione dei bad spikes
if (nargin > 4 & ~isempty(badSpikes))
   nbS = numel(badSpikes);
   if (nargin < 6 | isempty(badSpikesWidth))
       if (prod(size(spikesWidth)) == 1)
           badSpikesWidth = spikesWidth;
       else
           error('Inconsistent data for bad spikes');
       end
   end
   badSpikesWidth
   if (prod(size(badSpikesWidth)) == 1)
      badSpikesWidth = badSpikesWidth*ones(size(badSpikes));
   end
   rightEnds = badSpikes + badSpikesWidth;
   leftEnds  = badSpikes - badSpikesWidth;
% controllo sulla sovrapposizione
   overlaps  = leftEnds(2:end) - rightEnds(1:end-1);
   overlapsInd = find(overlaps < 0);
   if (~isempty(overlapsInd))
       newEnds = floor( (rightEnds(overlapsInd) - leftEnds(overlapsInd+1)) / 2 );
       rightEnds(overlapsInd)  = newEnds;
       leftEnds(overlapsInd+1) = newEnds;
   end
   % controllo su compatibilita' delle distanze
   leftEnds(1) = max( leftEnds(1), 1 );
   rightEnds(end) = min( rightEnds(end), nY );
% Calcolo dei valori di rimpiazzo
   newValues = (Y(leftEnds-1,:) + Y(rightEnds+1,:)) / 2; % (2 * nMS);
   for k = 1:length(badSpikes)
       for j = leftEnds(k):rightEnds(k)
           Y(j, :) = newValues(k, :);
       end
   end
end
% eventuale rimozione dei bad MZ
if (nargin > 6 & ~isempty(badMZ))
   Y(:, badMZ) = 0;
end
% allineamento
nS = size(spikes,1);
if (prod(size(spikesWidth)) == 1)
   spikesWidth = spikesWidth*ones(size(spikes));
end
deltad = diff(spikes);
[dref,drefpos] = max(deltad);     % reference distance
rightEnds = spikes - spikesWidth;
leftEnds  = spikes + spikesWidth;
rightEnds = rightEnds(2:nS);
leftEnds  = leftEnds(1:nS-1);
addPoints = dref - deltad;
newRightEnds = rightEnds + cumsum(addPoints);
newLeftEnds  = leftEnds + [0;cumsum(addPoints(1:end-1))];
newSpikes    = spikes + [0;cumsum(addPoints)];
nYS = nY + sum(addPoints);
```

```
YS  = zeros(nYS,nMS);
MZS = zeros(nYS,1);
YS(1:leftEnds(1)-1,:) = Y(1:leftEnds(1)-1,:);
MZS(1:leftEnds(1)-1)  = MZ(1:leftEnds(1)-1);
oldX = [1:nY]';
np   = length(addPoints);
for i = 1:np
    n0      = rightEnds(i) - leftEnds(i);
    oldY    = Y(leftEnds(i):rightEnds(i),:);
    oldMZ   = MZ(leftEnds(i):rightEnds(i));
    posOldY = [2:n0+1]';
    n = n0;
    if (addPoints(i))
        nPoints = addPoints(i);
        while( nPoints )
            m     = min(nPoints,n);
            newY  = [oldY(1,:); zeros(m+n,nMS)];
            newMZ = [oldMZ(1); zeros(m+n,1)];
            m1      = floor((n-m)/2);
            newInd  = [1:m1, m1+2:2:2*m+m1, 2*m+m1+1:n+m]'+1;
            posOldY = newInd(posOldY-1);
            newY(newInd,:) = oldY(2:end,:);
            newMZ(newInd)  = oldMZ(2:end);
            newY(m1+2:2:2*m+m1,:) = (oldY(m1+1:m+m1,:) + oldY(m1+2:m+m1+1,:)) / 2;
            newMZ(m1+2:2:2*m+m1)  = (oldMZ(m1+1:m+m1) + oldMZ(m1+2:m+m1+1)) / 2;
            nPoints = nPoints - m;
            n     = n + m;
            oldY  = newY;
            oldMZ = newMZ;
        end
    else
```

```
        newY  = oldY;
        newMZ = oldMZ;
    end
    if (posOldY(end) ~= n0+addPoints(i)+1)
         error(’Wrong number of added points’);
    end
    newX = [1:posOldY(end)]’; % questa dovrebbe non servire
    oldX(leftEnds(i)+1:rightEnds(i))     = newLeftEnds(i) + posOldY - 1;
    MZS(newLeftEnds(i):newRightEnds(i))  = newMZ;
    YS(newLeftEnds(i):newRightEnds(i),:) = newY;
    if (i < np)
       oldX(rightEnds(i)+1:leftEnds(i+1))         = newRightEnds(i)+1:newLeftEnds(i+1);
       MZS(newRightEnds(i)+1:newLeftEnds(i+1)-1)  = MZ(rightEnds(i)+1:leftEnds(i+1)-1);
       YS(newRightEnds(i)+1:newLeftEnds(i+1)-1,:) = Y(rightEnds(i)+1:leftEnds(i+1)-1,:);
    end
end
YS(newRightEnds(end)+1:nYS,:) = Y(rightEnds(end)+1:nY,:);
MZS(newRightEnds(end)+1:nYS)  = MZ(rightEnds(end)+1:nY);
oldX(rightEnds(end)+1:nY)     = newRightEnds(end)+1:nYS;
newX = [1:nYS]’;
```

Appendix B

```
function [pr,xa,hpr1,ntc,rapc,rapc2] = matty(dati,d)
if ( nargin < 1 | nargin > 2 )
error(’call is: [ac,nc] = autocorr(x,np)’);
end;
x=dati;
np=length(x);
x2=[0:1:np];
[n,m] = size(x);
```

```
xac=xcorr(x); % acf symm
xa=xac(m:2*m-1); % acf da zero
pr = xa./xa(1);  % normalizzazione
h1=pr(1:d:m-1);  % max della acf ad interdistanza d
nn=ceil(length(h1));
hpr1=h1;    % considera i primi 10 valori
nt=ceil(np/d); % numero totale di componenti ideali
nt4=ceil(nt/2);
hpr=hpr1*(nt)/(nt-1);  % correzione acf mac per numero di componenti
rac=(2-(4-4*hpr(2)*hpr(2))^0.5)/(2*hpr(2)); %ratio t 2t
rap=real(rac);
nttc=2*([1:nt4-1]./(1-hpr(3:2:(nt)))); % numero componenti calcolato su 4 massimi acf
ntc=mean(nttc(nt4-3:nt4-1));
racc=(2-(4-4*hpr(4)*hpr(4)/(hpr(5)*hpr(5)))^0.5)/(2*hpr(4)/hpr(5));
rapc=real(racc);
racc2=(2-(4-4*hpr(2)*hpr(2)/(hpr(3)*hpr(3)))^0.5)/(2*hpr(2)/hpr(3));
rapc2=real(racc2);
pr=pr';
```

List of Figures

2.1 First region of EACF plot 8
2.2 EACF plot with deterministic peaks (indicated by arrows) 8
2.3 GC-MS chromatogram of a not linearized mixture and the relative linearized one [31] 17
2.4 Simulated chromatogram and the relative $EACVF_{tot}$ with a R value equal to 2 [31] 22

3.1 Organic fraction of Particulate Matter [75] 27
3.2 BF_3/Buthanol Derivatization Procedure 31
3.3 BSTFA Derivatization Procedure 32

4.1 GC-MS chromatograms (SIM and TIC) and relative $EACVF_{tot}$ plots of a Fuel sample [32] 37
4.2 GC-MS chromatogram and relative $EACVF_{tot}$ plot of a *Mimosa* flower extract [32] 38
4.3 GC-MS chromatogram of the Sample **MI-17** [98] 40
4.4 $EACVF_{tot}$ plot of the Sample **MI-17** [98] 41
4.5 GC-MS chromatogram of a PM sample (BF_3/BuOH derivatization procedure) [129] 46
4.6 GC-MS chromatogram of a PM sample (BSTFA derivatization procedure) [129] . 47
4.7 GC-MS chromatogram (n-alkanoic acids) of $PM_{2.5}$ rural sample [132] 48
4.8 $EACVF_{tot}$ plot of n-alkanoic acids homologous series in a $PM_{2.5}$ rural sample [132] 49

List of Tables

4.1 Parameters of real samples [32] 36
4.2 Comparison between CPI values calculated through the traditional method and the $EACVF_{tot}$ one [98] 42
4.3 Concentrations (reported in ng/m^3) of the target dicarboxylic acids measured on two experimental PM samples after derivatization [129] 47
4.4 Parameters of a rural $PM_{2,5}$ sample (n-alkanoic acids series) [132] 49

PAPER I

Signal Processing of GC-MS Data of Complex Environmental Samples: Characterization of Homologous Series

Available online at www.sciencedirect.com
ScienceDirect
Analytica Chimica Acta 594 (2007) 128–138
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Signal processing of GC–MS data of complex environmental samples: Characterization of homologous series

Maria Chiara Pietrogrande *, Mattia Mercuriali, Luisa Pasti

Department of Chemistry, University of Ferrara, Via L. Borsari, 46, I-44100 Ferrara, Italy

Received 1 February 2007; received in revised form 7 May 2007; accepted 13 May 2007
Available online 21 May 2007

Abstract

Identification and characterization of homologous series by GC–MS analysis provide very relevant information on organic compounds in complex mixtures. A chemometric approach, based on the study of the autocovariance function, $\mathrm{EACVF_{tot}}$, is described as a suitable tool for extracting molecular–structural information from the GC signal, in particular for identifying the presence of homologous series and quantifying the number of their terms. A data pre-processing procedure is introduced to transform the time axis in order to display a strictly homogenous retention pattern: *n*-alkanes are used as external standard to stretch or shrink the original chromatogram in order to build up a linear GC retention scale. This addition can be regarded as a further step in the direction of a signal processing procedure for achieving a systematic characterization of complex mixture from experimental chromatograms. The $\mathrm{EACVF_{tot}}$ was computed on the linearized chromatogram: if the sample presents terms of homologous series, the $\mathrm{EACVF_{tot}}$ plot shows well-defined deterministic peaks at repeated constant interdistances. By comparison with standard references, the presence of such peaks is diagnostic for the presence of the ordered series, their position can be related to the chemical structure of the compounds, their height is the basis for estimating the number of terms in the series. The power of the procedure can be magnified by studying SIM chromatograms acquired at specific *m*/*z* values characteristic of the compounds of interest: the $\mathrm{EACVF_{tot}}$ on these selective signals makes it possible to confirm the results obtained from an unknown mixture and check their reliability.

The procedure was validated on standard mixtures of known composition and applied to an unknown gas oil sample. In particular, the paper focuses on the study of two specific classes of compounds: *n*-alkanes and oxygen-containing compounds, since their identification provides information useful for characterizing the chemical composition of many samples of different origin. The robustness of the method was tested in experimental chromatograms obtained under unfavorable conditions: chromatograms acquired in non-optimal temperature program conditions and chromatographic data affected by signal noise.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Multicomponent mixtures; Homologous series; Time axis linearization; Hydrocarbons

---

1. Introduction

Real world samples submitted to chromatographic analysis are usually very complex matrices made up of thousands of components which display a wide range of concentrations. It is practically impossible to achieve chromatographic characterization of all the compounds in a sample in a single separation step. Indeed, even when hyphenation with mass spectrometry is employed, it is only possible to identify and quantify a limited number of compounds belonging to different group types [1,2]. For this reason, methods that separate organics into compound classes (as opposed to individual compounds) are often able to characterize a large fraction of the organic components [3,4]. The chromatograms thus obtained are complex signals characterized by strong peak overlapping (in particular for 1D separations) and formed by an extensive amount of data (in particular for 2D and hyphenated separations) [5,6]. Even if the mixture is not sufficiently well separated in any one dimension, a great deal of chemical information regarding the mixture can be obtained from the chromatographic data by applying spe-

---

Abbreviations: ACVF, autocovariance function; EACVF, experimental autocovariance function; EACF, experimental autocorrelation function; SC, single component

* Corresponding author. Tel.: +39 0532291152.

E-mail address: mpc@unife.it (M.C. Pietrogrande).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.05.020

cific signal processing procedures devoted to extracting all the analytical information on separation and sample [7–11].

This paper presents a mathematical–statistical approach developed to single out information from a complex 1D chromatogram. It is a chemometric approach based on the study of the autocovariance function (ACVF) which has proved itself a helpful tool in decoding complex chromatograms, i.e., extracting information on the mixture—number of components, abundance distribution – and separation—separation performance, retention pattern [12–22]. In particular, it is powerful in magnifying any ordered retention patterns present in the chromatogram, and singling them out from the "disordered forest" of random peaks [16–18]. Such an order can be related to chemical structure of the sample components under a strictly homogenous retention pattern.

The present paper describes a linearization algorithm using *n*-alkanes as external standard to stretch or shrink the original chromatogram in order to build up a linear GC retention scale. The introduction of this data pre-processing algorithm can be regarded as a further step in the direction of a signal processing procedure for achieving a systematic characterization of complex mixture from experimental chromatograms. The study is focused on the characterization of organics present in the sample in terms of structural group analysis, e.g., chemical class identification by homologous class, polarity, or functionality. This provides useful information for the characterization of many different types of samples. For example, the presence of organics in complex environmental samples can be quite useful in understanding source contributions (e.g., biogenic versus anthropogenic), sample evolution and sample environmental fate as related to health impact [23–28].

In particular, this paper is devoted to the study of two specific classes of compounds: *n*-alkanes and oxygen-containing compounds, whose identification provides relevant information in evaluating the contribution that local emission sources, naturally produced aerosols, transportation, and industrial activities make to local air quality [26,29,30]. A petrochemical matrix was studied, since it represents a sample that can be characterized by the structural group analysis: detailed structural information in terms of paraffinic, olefinic, naphthenic and aromatic carbon and hetero-compounds is relevant to characterize crude oil and oil products, and it can be used in monitoring petroleum industrial processing steps and in investigating the presence of petrochemical pollutants in air, water and soil. It could also be advantageous in analyzing the space-, time- or particle size-dependent variability of the chemical fingerprints as well as for source characterization [23–25].

2. Theory

2.1. Autocovariance function method

The chemometric approach studies the autocovariance function that can be directly computed from the experimental chromatogram acquired in digitized form (Experimental ACVF, EACVF), using the following expression [12]:

$$\mathrm{EACVF}(\Delta t) = \frac{1}{M}\sum_{j=1}^{M-k}(Y_j - \hat{Y})(Y_{j+k} - \hat{Y}) \qquad k = 0, 1, 2, \ldots M-1 \tag{1}$$

where Y_j is the digitized chromatogram signal, $\hat{Y}$ its mean value and M the truncation point in the EACVF computation.

Theoretical models have been developed to describe ACVF (theoretical ACVF, TACVF) in terms of the complex chromatogram parameters, i.e., number of single components (SC), m_{tot}, SC peak standard deviation, σ, peak abundance distribution [12].

The most general case is a disordered multicomponent chromatogram containing m_{tot} SCs displaying a Poissonian retention pattern. Assuming chromatographic peaks of Gaussian shape with constant width, i.e., optimized programmed temperature conditions, the value of $\mathrm{EACVF}_{\mathrm{tot}}$ at the origin ($\Delta t = 0$) is described by a simple equation from which the number of SCs of the chromatogram, m_{tot} can be estimated as [15]:

$$m_{\mathrm{tot}} = \frac{A^2_{\mathrm{T,tot}}(\sigma^2_{\mathrm{M,tot}}/a^2_{\mathrm{M,tot}} + 1)}{\mathrm{EACVF}_{\mathrm{tot}}(0)d_{\mathrm{h/2}}2.129X} \tag{2}$$

where $A_{\mathrm{T,tot}}$ is the total area of the chromatogram, $d_{\mathrm{h/2}}$ the half height width of the $\mathrm{EACVF}_{\mathrm{tot}}$ peak—it can be simply related to the mean peak standard deviation, σ —, X is the total chromatogram time range, $\sigma^2_{\mathrm{M,tot}}/a^2_{\mathrm{M,tot}}$ the peak maximum dispersion ratio derived from the mean, $a_{\mathrm{M,tot}}$, and the variance, $\sigma^2_{M,\mathrm{tot}}$, of peak maxima computed from the observed peak maxima in the chromatogram [15].

Another model is an ordered chromatogram formed by a sequence of $n_{\max}$ SC peaks where the retention time of the n-th term is described by:

$$t_{\mathrm{R}}(n) = c + bn, \quad n = 0, 1, 2, 3, \ldots, n_{\max} \tag{3}$$

This is the case of terms of a homologous series submitted to GC analysis under optimized, linearized temperature programming conditions: c represents the contribution of a specific functional group to the overall retention, b the retention increment between terms of the homologous series, e.g., the $\mathrm{CH_2}$ retention time increment [31,32].

For the ordered retention pattern (indicated by the subscript O), the $\mathrm{TACVF_O}$, and therefore the $\mathrm{EACVF_O}$, plot displays well-defined Gaussian peaks located at interdistances bk, corresponding to repeated interdistances between terms of the ordered series (Eq. (3)). These peaks are diagnostic and identify the presence of terms of the homologous series in the sample. The value of $\mathrm{EACVF_O}$ at the repeated interdistances ($\Delta t = bk$) can be used to estimate the number of SCs belonging to the ordered series, $n_{\max}$, according to the following equation [22]:

$$n_{\max} - k = \frac{A^2_{\mathrm{T,O}}(\sigma^2_{\mathrm{M,O}}/a^2_{\mathrm{M,O}} + 1)}{\mathrm{EACVF_O}(bk)d_{\mathrm{h/2}}2.129X} \tag{4}$$

In the most general case, a multicomponent mixture is formed by combining SCs with uncorrelated chemical structures, m_{P}, displaying Poissonian retention pattern, with $n_{\max}$ SCs belonging to homologous series yielding an ordered pattern. This means that the total number of SCs present in the mixture is $m_{\text{tot}} = m_{\text{P}} + n_{\max}$ and the complex chromatogram is the combination of a random retention pattern yielded by m_{P} SCs (Eq. (2)) and an ordered component due to the $n_{\max}$ terms of the homologous series (Eq. (3)) [22]. The $\text{EACVF}_{\text{tot}}$ computed on the complex chromatogram may be handled as the superimposition of a Poissonian EACVF_{P} and an ordered EACVF_{O}: this is the consequence of the variance additivity for independent variables, since the Poissonian and ordered parts are not correlated. Therefore, a combination of Eqs. (2)–(4) can be applied to estimate the mixture composition [22].

Eq. (2) can be used to estimate the total number of components m_{tot} in the mixture. In the $\text{EACVF}_{\text{tot}}$ plot, the presence of deterministic peaks at repeated interdistances ($\Delta t_{\text{R}} = bk$) is mainly due to the contribution of the ordered component (Eq. (4)) since EACVF_P assumes values close to 0 for $\Delta t \geq 4\sigma$. Such deterministic peaks indicate that the sample contains terms of homologous series and Eq. (4) can be used to evaluate the $n_{\max}$ value from $\text{EACVF}_{\text{tot}}$ (bk): the parameters $A_{\text{M,O}}$ and $\sigma^2_{\text{M,O}}/a^2_{\text{M,O}}$ in Eq. (4) — concerning the ordered component only — cannot be experimentally determined and must be approximated by $A_{\text{M,tot}}$ and $\sigma^2_{\text{M,tot}}/a^2_{\text{M,tot}}$ values computed on the whole chromatogram. This assumption is possible if the two following equations are true:

$$\sigma^2_{\text{M,tot}}/a^2_{\text{M,tot}} \approx \sigma^2_{\text{M,O}}/a^2_{\text{M,O}} \tag{5}$$

$$A_{m,\text{tot}} = \frac{A_{\text{T,tot}}}{m_{\text{tot}}} \approx A_{m,\text{O}} = \frac{A_{n_{\max},\text{O}}}{n_{\max}} \tag{6}$$

The first condition (Eq. (5)) assumes that the SCs belonging to a given homologous series display the same peak maximum dispersion ratio, $\sigma^2_{\text{M}}/a^2_{\text{M}}$, as that of all the SCs in the mixture: it is usually met in real samples, since the SC abundances generally follow the most likely exponential distribution, yielding: $\sigma^2_{\text{M}}/a^2_{\text{M}} \approx 1$.

Eq. (6) concerns the average peak areas A_m: the hypothesis assumes that the $A_{n_{\max},\text{O}}$ value, computed on the $n_{\max}$ components of the ordered series, is the same as $A_{m,\text{tot}}$ determined on the total chromatogram. This condition is not usually met in the practice: in general one homologous series is predominant or in trace versus most of the other SCs [1,23,25].

In the simplest case, when both the Eqs. (5) and (6) are strictly true, for $k = 1$ a combination of Eqs. (2) and (4) yields a simple equation able to estimate the $n_{\max}$ value:

$$n_{\max} = m_{\text{tot}} \frac{\text{EACVF}_{\text{tot}}(b)}{\text{EACVF}_{\text{tot}}(0)} + 1 \tag{7}$$

Conditions: Eqs. (5) and (6); $b \geq 4\sigma$.

In the most general case, when only Eq. (6) holds true, by combining Eqs. (2) and (4), one can obtain:

$$n_{\max} = m_{\text{tot}} \times \frac{\text{EACVF}_{\text{tot}}(b)}{\text{EACVF}_{\text{tot}}(0)} \times \frac{A^2_{m,\text{tot}}}{A^2_{n_{\max},\text{O}}} + 1 \tag{8}$$

Conditions: Eq. (6); $b \geq 4\sigma$.

Since the two quantities $A_{m,\text{tot}}$ and $A_{m,\text{O}}$ are not experimentally accessible, from $\text{EACVF}_{\text{tot}}$ it is possible to estimate an "apparent" $n_{\max}$, given by:

$$n_{\max,\text{ap}} = n_{\max} \times \frac{A^2_{m,\text{O}}}{A^2_{m,\text{tot}}} \tag{9}$$

It is the apparent $n_{\max}$ value computed under the hypothesis of the same mean peak area displayed by the terms of the ordered series and all the components. It represents a relative estimation of $n_{\max}$, depending on the relative mean area of the ordered components, $A_{m,\text{O}}$, compared to the mean area of the whole mixture, $A_{m,\text{tot}}$.

A simplified procedure can be obtained by computing the autocovariance function EACF_{tot} (Δt_{R}), i.e., the ratio $\text{EACVF}_{\text{tot}}(\Delta t_{\text{R}})/\text{EACVF}_{\text{tot}}(0)$: the EACF_{tot} (b) value can be simply related to $n_{\max}$ (Eq. (7)) or to $n_{\max,\text{ap}}$ (Eq. (8)) in the most general case:

$$\text{EACF}(b) = \frac{\text{EACVF}_{\text{tot}}(b)}{\text{EACVF}_{\text{tot}}(0)} = \frac{n_{\max,\text{ap}}}{m_{\text{tot}}} - 1 \tag{10}$$

Therefore, the study of $\text{EACVF}_{\text{tot}}$ makes it possible to estimate the complexity of the whole mixture (m_{tot}, Eq. (2)) and identify the presence, and the specific contribution, of the components belonging to the ordered series ($n_{\max}$ or $n_{\max,\text{ap}}$, Eqs. (7) and (8)), singling it out from the disordered pattern. Note that the described approach is general, since $\text{EACVF}_{\text{tot}}$ (Δt_{R}) is a quantity which can be numerically computed from the chromatogram without any *a priori* assumption or theoretical model.

2.2. Linearization procedure

The study of $\text{EACVF}_{\text{tot}}$ to identify the sample chemical composition and extract structural information regarding the mixture components from the GC signal is based on a strict linear relation between the retention time $t_{\text{R}}(n)$ and number of repeated units n within a homologous series (Eq. (3)). This is true under linear temperature-programmed GC conditions, as confirmed by both experimental evidence and theoretical studies based on retention thermodynamics [18,31,32]. However, the strictly homogenous retention pattern yielding constant retention increments between subsequent terms of homologous series is difficult to be achieved in the practice because of experimental limitations, i.e., the not strictly linear temperature-programmed GC runs, poor reproducibility in flow rate or temperature, variations in injection-timing and temperature program rate [31,32].

Therefore, in order to usefully apply the $\text{EACVF}_{\text{tot}}$ procedure, a data handling algorithm is required to linearize experimental chromatograms prior to $\text{EACVF}_{\text{tot}}$ computation. If $Y(x)$ represents the chromatographic signal, where x is the retention time, the time axis is transformed into a new scale by using a function $z = g(x)$ to relate the original time axis to the new z axis. In this transformation the total signal area must be preserved: this means transforming the signal $Y_1(x)$ at a given x position in the original chromatogram into the corresponding $Y_2(z)$ value at the z position in the transformed chromatogram so that the following condition is fulfilled:

$$Y_1(x)\mathrm{d}x = Y_2(z)\mathrm{d}z \tag{11}$$

Instead of a continuous function $z = g\ (x)$, the present paper proposes an empirical transformation procedure based on an equidistant retention position between the subsequent terms of *n*-alkane homologous series. This means that the applied transformation has the property:

$$Y_1(x)\Delta x = Y_2(z)\Delta z \tag{11a}$$

over finite Δx and Δz regions of the chromatograms between subsequent terms of a homologous series, i.e., CH_2 addition retention increment.

The use of *n*-alkanes as external standard to build up a GC retention scale is very common in GC: in fact, the *n*-alkanes may act as the flexible "mile-stone" system of the chromatogram, and the relative position of the analyte compounds can be referred to them [33,34]. An *n*-alkane reference mixture containing the terms displaying retention values in the same range as the sample was analyzed under the same temperature-programmed GC conditions used for the unknown sample. Within a given threshold distance, the *n*-alkane reference peaks are matched to the nearest peaks in the sample chromatogram. The sample signal was divided into many regions corresponding to the distance $\Delta x = \Delta t_R$ between subsequent terms of *n*-hydrocarbons; a Δx value (usually the average of experimental Δx values) was selected as constant Δz retention increment in the new scale. Each inter-peak region is taken individually and it is stretched, or shrunk, to force each Δx interdistance to the constant Δz value.

3. Experimental

3.1. Organic mixtures

The standard mixtures contained known amount of organic compounds: C_6–C_{32} *n*-hydrocarbons, and some organic compounds with uncorrelated molecular structures (alcohols, ketones, esters and aromatics with carbon atomic numbers ranging from 3 to 11) were purchased from Aldrich and from Supelco (Milan, Italy) (99% min). The standard solutions of organic acids contained 16 *n*-alkanoic acids (from C_8 to C_{23}), benzene carboxylic acids (1,2 and 1,3 benzenedicarboxylic, 1,2,3 and 1,3,5 benzenetricarboxylic acids) and amino acids (20 small molecules); they were purchased from Aldrich (99% min). The mixtures were prepared by mixing proper concentrations of standard compounds so that all the terms of the homologous series display nearly the same mean peak area values (Eq. (6)) and the same peak maxima dispersion ratios (Eq. (5)).

The unknown sample studied was an ASTM D2887 Reference Oil (C_6–C_{44}, b.p. 115–475 °C) that is the basis for standard test method for boiling range distribution of petroleum fractions by GC. It was delivered from Supelco (Milan, Italy).

3.2. Derivatization procedure of organic acids

Before GC analysis, the carboxylic acids were submitted to chemical derivatization with MTBSTFA (*N*,*N*-Methyl-*tert*-butyl(dimethyl-silyl)trifluoroacetamide (Interchim, France): 30 μl of MTBSTFA was added as reactant to the sample in pyridine (10 μl) as previously described [21]. *N*,*N*-Methyl-*tert*-butyl(dimethyl-silyl)trifluoroacetamide (MTBSTFA) and pyridine were obtained from Interchim (France) and from Fluka (France), respectively.

3.3. Instrumentation

The GC–MS analyses were performed on a Mega Series 5160 gas chromatograph (Fisons Instruments, Milan, Italy) coupled with a QMD1000 quadrupole mass spectrometer (Fisons Instruments, Milan, Italy). The column used was a DB-5 column (L = 30 m, I.D. 0.25 mm, d_f 0.25 μm) (J&W Scientific, Rancho Cordova, CA, USA). The analyses were performed under different programming conditions according to the specific sample analyzed. The carrier gas was helium at a flow rate of 1.2 ml min^{-1}. Split conditions (1:400 split ratio) were used for injection (injection temperature: 200 or 300 °C; injected sample: 1 μl of mixture). The mass spectrometer operated in EI mode (positive ion, 70 eV): mass spectra were acquired with repetitive scanning from 40 to 400 *m*/*z* in 1 s.

To analyze the carboxylic acids, the split/splitless injector was operated at 300 °C (mean split ratio: 1:20). The source was heated to 270 °C and helium was used as carrier gas [34]. Temperature-programmed analysis was performed increasing from 100 to 280 °C, at a rate of 3 °C min^{-1} [34].

3.4. Computation

All the programs are written in Fortran and run on a 2 GHz (512 RAM), Pentium III personal computer.

3.4.1. EACVF calculation

The EACVF was numerically calculated from the digitized chromatogram, according to Eq. (1) [12]. The developed algorithm concerns the calculation of some separation parameters reported in Eqs. (2) and (4) [22]: the total area of the chromatogram, A_T, was computed by numerical integration. The peak maxima were detected in the chromatogram by using an algorithm that compares five successive points and a threshold level to filter out the noise. The average peak maximum abundance a_M, and its standard deviation σ_M^2, were computed from these values. According to Eq. (2), the number of SCs, m_{tot}, can be estimated as $m_{tot} \pm \sqrt{m_{tot}}$ because of the Poissonian character of this variable.

3.4.2. Linearization procedure

A temperature GC analysis program was chosen to yield nearly linear retention behavior for *n*-alkane reference mixture containing the terms displaying retention values close to the sample: the *n*-alkane reference and the sample were analyzed under the same GC conditions.

The linearization algorithm identifies the *n*-alkane peaks in the reference mixture and matches them in the sample chromatogram by comparing the appearance surrounding their retention time, i.e., ±5 points. If an *n*-alkane peak is not matched, the retention value of the reference chromatogram is assumed in the sample chromatogram. Then, on the basis of the identified *n*-alkane peaks, the algorithm divides the sample chromatogram time axis into a number of regions corresponding to retention increment $\Delta x = \Delta t_R$ between subsequent *n*-alkane terms. A Δx value (usually the average of experimental Δx values) is selected as constant Δz retention increment in the new scale. The linearization algorithm works on each retention interval individually: to force each Δx interdistance to the constant Δz value, each Δx interval is compressed, by deleting signal points, or expanded, by adding signal data, to reach the same Δz interval. A proper data handling algorithm was developed for the addition, or deletion, of data points in order to preserve the total area of the original chromatogram in the re-scaled signal (Eq. (11a)).

4. Results and discussion

The reliability of the proposed method was verified on standard mixtures containing homologous series terms plus some uncorrelated compounds of known abundance and distribution. The unknown sample selected to test the method applicability was a petrochemical sample, since it represents a complex sample where the structural group analysis is quite relevant, able to provide information for the sample composition characterization.

4.1. Validation using standard mixtures

4.1.1. Analysis of n-alkanes

The method was validated by using standard mixtures of *n*-alkanes. Structural information on these compounds characterize the chemical nature of petrochemical matrices, e.g., paraffinic, olefinic, naphthenic and aromatic carbon [5–7,24]. Such information is relevant in monitoring the industrial processes used by the petroleum industry and in analyzing petrochemical-derived air, water and soil pollution [23–28].

The standard mixture studied contained six subsequent terms of *n*-alkane series (C_7–C_{12}) in addition to 14 organic compounds with uncorrelated structures. It was analyzed under a temperature program that started at 30 °C for 3 min and then increased to 80 °C at 5° min^{-1}, as this was found to approach the optimal temperature programming conditions.

The original chromatogram (Fig. 1a) was submitted to the linearization procedure using the signal obtained from the reference C_7–C_{12} *n*-alkanes analyzed under the same experimental conditions (inset in Fig. 1a): a value $\Delta t_R = 2.5$ min was selected as constant retention increment between subsequent terms of the series (*b* in Eq. (2), arrows in the chromatograms in Fig. 1b). The EACVF$_{tot}$ was computed on the linearized chromatogram (Fig. 1b). The number of the mixture components, m_{tot}, was estimated from the EACVF$_{tot}$(0) by using Eq. (2): the obtained value $m_{tot} = 19 \pm 4$ fully agreed with the real experimental value $m_{tot} = 20$. The EACVF$_{tot}$ plot (Fig. 2a, plain line) clearly shows deterministic peaks at repeated positions, i.e., 2.5, 5.0 min. These peaks are diagnostic for the presence of an ordered series, singling it out from the disordered pattern of peaks present in the chromatogram (Fig. 1a). The presence of the homologous series can be inferred by the coincidence with EACVF$_{tot}$ deterministic peaks obtained from the linearized chromatogram of the reference mixture (Fig. 2a, dotted line). From the EACVF$_{tot}$ computed at $\Delta t_R = 2.5$ min it is possible to estimate the number of terms of the homologous series, n_{max}, by using Eq. (7): the obtained value $n_{max} = 6$ is exactly the real experimental value. It must be underlined that Eq. (7) can be correctly applied in this case: proper concentrations of standard compounds were added to prepare a mixture fulfilling the conditions on SC abundance distribution described by Eqs. (5) and (6).

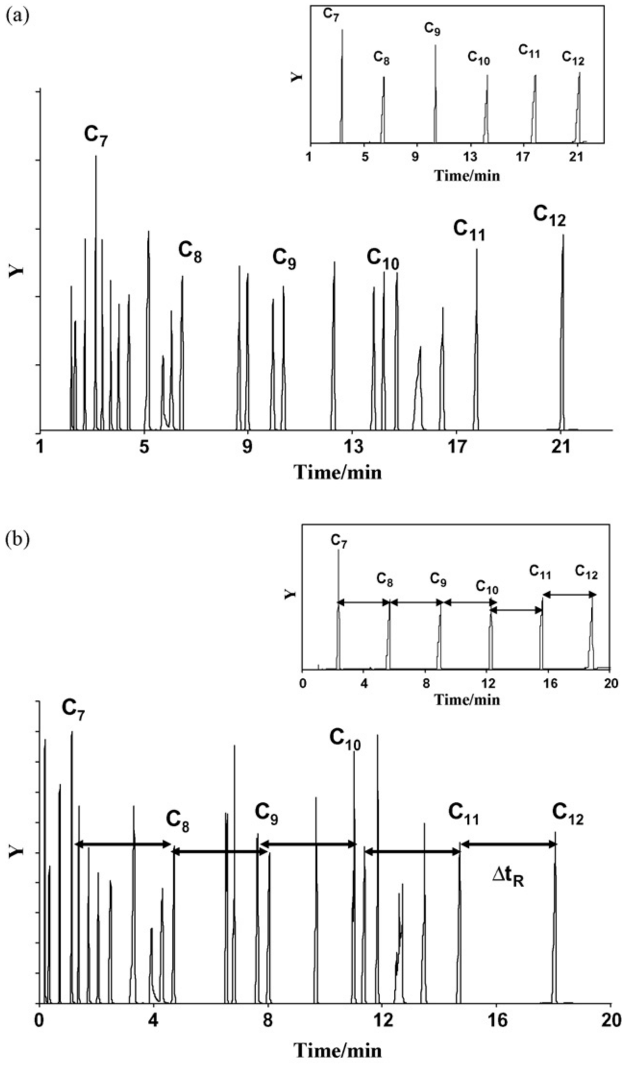


Fig. 1. GC–MS TIC chromatogram of a standard mixture containing six subsequent terms of *n*-alkane series (C_7–C_{12}) in addition to 14 organic compounds with uncorrelated structures. Temperature program: 30 °C for 3 min, an increase to 80 °C at 5° min^{-1}. (a) Original chromatograms of the standard mixture and the reference C_7–C_{12} *n*-alkanes (in the inset) and (b) linearized chromatograms of the standard mixture and the reference C_7–C_{12} *n*-alkanes (in the inset): the arrows indicate the $\Delta t_R = 2.5$ min value selected as the constant retention increment.

The effect of the linearization procedure can be shown by comparing the EACVF$_{tot}$ (Fig. 2a, plain line) obtained from

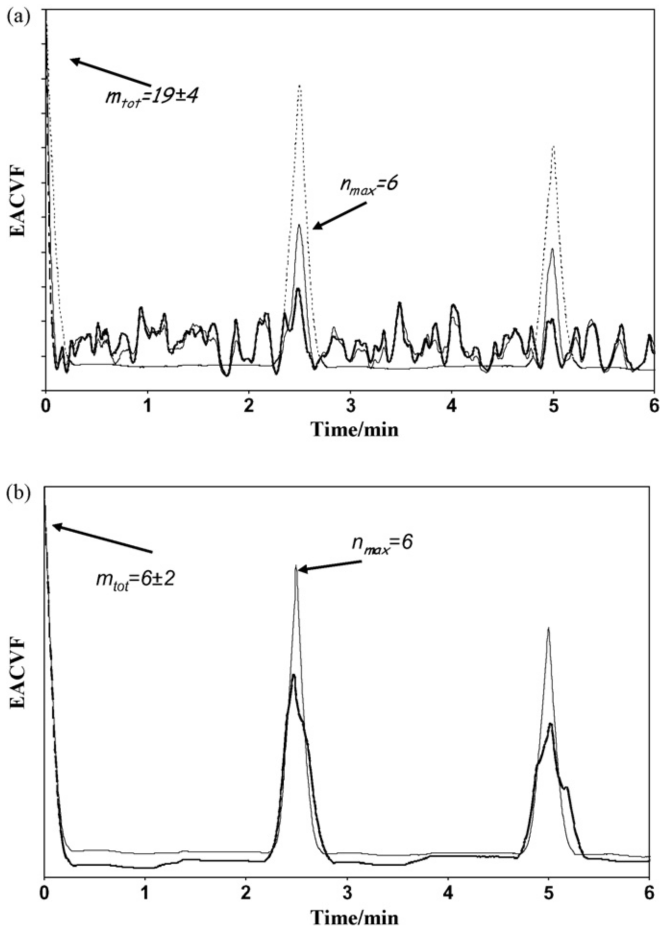


Fig. 2. EACVF$_{tot}$ plot computed on the chromatograms (Fig. 1). (a) EACVF$_{tot}$ plot computed on the chromatogram of a standard mixture containing 20 organic compounds. Plain line: linearized chromatogram; bold line: original chromatogram under temperature programming conditions close to linearity; dotted line: linearized chromatogram of the reference mixture and (b) EACVF$_{tot}$ plot computed on the chromatogram of the reference C_7–C_{12} *n*-alkanes. Plain line: linearized chromatogram; bold line: original chromatogram under temperature programming conditions close to linearity.

the linearized chromatogram (Fig. 1b) to that (Fig. 2a, bold line) computed on the original chromatogram obtained under temperature programming conditions that are close to linearity (Fig. 1a). In this last EACVF$_{tot}$ plot, the deterministic peaks at repeated positions, i.e., 2.5, 5.0 min which are diagnostic for the presence of the homologous series – are still present, but lower thus making it difficult to identify them in the complex signal. These results underline the fact that signal linearization, performed using a mathematical procedure, is necessary to simply single out the retention repetitivities and therefore fully extract chemical information contained in the retention pattern.

4.1.2. Analysis of organic acids

The proposed procedure was tested on standard solutions of organic acids: the presence of oxygen-containing compounds (aldehydes, ketones, alcohols, and carboxylic acids) in environmental samples is diagnostic for secondary gas-phase and particle-phase organic matter resulting from photochemical conversion of primary biogenic and anthropogenic atmospheric compounds. As they undergo photochemical reactions, atmospheric organic matter becomes oxidized via reactions with several major oxidizing species in the atmosphere: the hydroxyl radical, the nitrate radical, and ozone [26,28–30].

Prior to GC injection, these polar organics must be converted into non-polar compounds that will then elute through the GC column. The studied mixture contained 16 *n*-alkanoic acids (C_8–C_{23} terms) and other mono-carboxylic acids with uncorrelated molecular structures (benzene carboxylic and amino acids). The MTBSTFA derivatization procedure was selected to yield derivatives suitable to GC analysis [21]. The TIC chromatogram of a standard mixture containing 50 organic compounds is reported in Fig. 3a. The signal was linearized on the basis of a reference *n*-alkanoic acid mixture (C_8–C_{23} terms) analyzed under the same analytical conditions (inset in Fig. 3a): a constant interdistance $\Delta t_R = 3$ min was selected as retention contribution for CH_2 increment in the series.

The EACVF$_{tot}$ was computed on the linearized chromatogram (Fig. 3b, bold line). The number of components, m_{tot}, was estimated from the EACVF$_{tot}$(0) (Eq. (2)): the obtained value was $m_{tot} = 49 \pm 7$ showing an excellent agreement with the effective number of components, 50, present in the standard mixture.

Information on the chemical composition of the sample can be extracted by a simple inspection of the EACVF$_{tot}$ plot (Fig. 3b, bold line): it clearly shows deterministic peaks at 3 min and repeated interdistances, diagnostic for the presence of a homologous series. Under the applied experimental conditions, comparison with the EACVF$_{tot}$ plot (Fig. 3b, plain line) computed on the *n*-alkanoic acid reference mixture (inset in Fig. 3a) identifies the series as *n*-alkanoic acids. From the EACVF$_{tot}$ value computed at $\Delta t_R = 3$ min the number of terms in the homologous series, n_{max}, can be estimated by using Eq. (7): the obtained value is $n_{max} = 16$, which corresponds exactly to the real number in the sample. In this case Eq. (7) can be properly applied since the standard mixture was properly prepared "*a priori*" to meet the assumptions on SC peak height distribution (Eqs. (5) and (6)). It must be underlined that the information on the presence and number of *n*-alkanoic acids can be simply extracted from the experimental MS-TIC chromatogram, without any time-consuming investigation on MS spectra of each peak.

In the case of a totally unknown mixture, the reliability of the results obtained with the EACVF$_{tot}$ may be checked by studying the SIM chromatogram acquired by selecting a specific mass fragment characteristic of the *n*-alkanoic acid derivatives. It has been found that the MS spectra of the *n*-alkanoic acid derivatives show a characteristic fragment at $m/z = 75$ which is due to the group $[OSi\text{-}(CH_3)_2]^+$ resulting from fragmentation of the silyl derivative [21].

The SIM chromatogram of the complex mixture was monitored at $m/z = 75$: the obtained signal (Fig. 3c) is simpler than the TIC chromatogram, since it mainly retains the signal of the *n*-alkanoic acids. In the case of unknown samples, such a chromatogram can prove helpful in verifying "*a posteriori*" the assumptions regarding the quantities $\sigma^2_{M,tot}/a^2_{M,tot}$ and

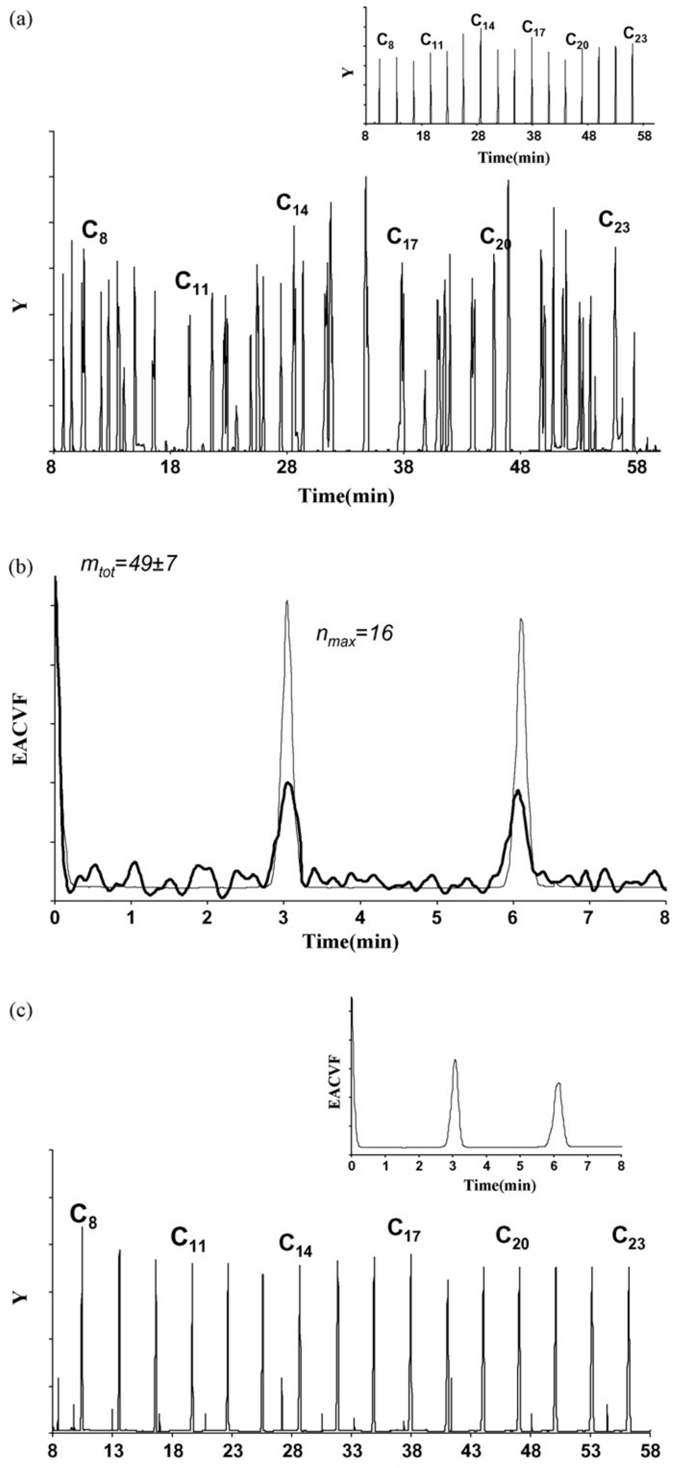


Fig. 3. GC–MS chromatograms of standard mixtures of organic acids. (a) Linearized GC–MS chromatogram of a standard mixture containing 50 organic acids; inset: chromatogram of a reference *n*-alkanoic acid mixture (C_8–C_{23} terms); (b) $\text{EACVF}_{\text{tot}}$ plots computed on the chromatograms. Bold line: $\text{EACVF}_{\text{tot}}$ on the chromatogram of organic acid mixture (Fig. 3a); solid line: $\text{EACVF}_{\text{tot}}$ on the chromatogram of the *n*-alkanoic acid reference mixture (inset in Fig. 3a) and (c) GC–MS SIM chromatogram of a standard mixture containing 50 organic acids at $m/z = 75$ mainly displaying *n*-alkanoic acid signal; inset: $\text{EACVF}_{\text{tot}}$ plot computed on the chromatogram.

$\sigma^2_{n_{\max},\text{O}}/a^2_{n_{\max},\text{O}}$ (Eq. (5)), $A_{m,\text{tot}}$ and $A_{n_{\max},\text{O}}$ (Eq. (6)) which serve as the basis for applying Eqs. (7) and (8). In fact, the TIC chromatogram represents the whole complex mixture (suffix tot) and the SIM signal selectively shows the ordered component (suffix O), assuming uniform response factors for the *n*-alkanoic acids in both acquisition modes. For the ordered component, the SC abundance distribution $\sigma^2_{n_{\max},\text{O}}/a^2_{n_{\max},\text{O}}$ can be computed from the SIM signal and compared to the corresponding value for the whole mixture,$\sigma^2_{\text{M,tot}}/a^2_{\text{M,tot}}$. Thus, the values obtained—$\sigma^2_{n_{\max},\text{O}}/a^2_{n_{\max},\text{O}} = 0.03$ and—indicate that both abundance distributions are described by the same uniform function and therefore the condition described by Eq. (5) is fulfilled. Otherwise, the parameter $A_{n_{\max},\text{O}}$ for an unknown sample cannot be estimated from the SIM signal; thus the condition $A_{m,\text{tot}} \approx A_{n_{\max},\text{O}}$ (Eq. (6)) is not experimentally verifiable and the relative quantity $n_{\text{max,ap}}$ can be estimated (Eq. (8)).

The $\text{EACVF}_{\text{tot}}$ calculated on the SIM chromatogram (inset in Fig. 3c) clearly shows the deterministic peaks at 3 min and multiple values diagnostic for *n*-alkanoic acids. The value $n_{\text{max,ap}} = 16$ is obtained from the $\text{EACVF}_{\text{tot}}$ value at $\Delta t = 3$ min and corresponds precisely to the value calculated from the TIC signal. The coincidence between the results independently obtained from the TIC and SIM chromatograms may be a cross-check of procedure reliability for unknown samples, i.e., the estimated $n_{\text{max,ap}}$ values coincide with the real $n_{\max}$ value.

4.2. Method robustness

The method was applied to experimental chromatograms obtained under unfavorable conditions in order to verify the robustness of the method, i.e., its power to overcome problems related to signal acquisition and to obtain reliable results. The first case concerns the effect of deviation from strictly linear retention conditions; the second study is devoted to the contribution of signal noise present in the chromatographic data.

4.2.1. Non-linear chromatographic retention

The relevance of the chromatogram linearization preliminary to the $\text{EACVF}_{\text{tot}}$ study has previously been discussed. An incomplete linearization of the time axis is a usual feature in experimental signals, since some slight difference in Δt_{R} between subsequent terms of the homologous series is often displayed, even if operating conditions are close to the optimal temperature program.

To test the robustness of the $\text{EACVF}_{\text{tot}}$ procedure in a moderate deviation from retention linearity, $\text{EACVF}_{\text{tot}}$ was computed on the original signal from the reference mixture containing C_7–C_{12} *n*-alkanes (temperature programming conditions close to the linearity, inset in Fig. 1a) and the obtained plot (Fig. 2b, bold line) was compared with the $\text{EACVF}_{\text{tot}}$ obtained from the linearized chromatogram (Fig. 2b, plain line). At repeated positions, i.e., 2.5, 5.0 min, the deterministic peaks, which are diagnostic for the presence of the homologous series, are still present, but their maximum values are significantly lower which leads to underestimation of the $n_{\max}$ value: a value $n_{\max} = 4$ (Eq. (7)) is estimated from Fig. 2b, $\text{EACVF}_{\text{tot}}$ computed at

$\Delta t_R = 2.5$ min and this value is lower than the experimental $n_{max} = 6$. However, the EACVF$_{tot}$ procedure also provides the means for overcoming this drawback. A close examination of the EACVF$_{tot}$ plot obtained from the non-linearized signal (Fig. 2b, bold line) shows a larger peak (larger $d_{1/2}$ value) than the one computed on the linearized signal (Fig. 2b, plain line). This discrepancy can be fully explained by close examination of Eq. (4), which lies at the basis of Eq. (7) for computing n_{max}: Eq. (4) contains the product EACVF$_O$ (bk) $d_{1/2}$ and therefore a decrease in the EACVF$_O$ (bk) value may be compensated by an increase in $d_{1/2}$ to obtain a correct estimation of n_{max}. For example, at $\Delta t_R = 2.5$ min, the EACVF$_{tot}$ peak computed on the original signal (Fig. 2b, bold line) shows a maximum value of 0.47 and $d_{1/2}$ of 0.26 min. The corresponding peak computed on the linearized signal (Fig. 2b, plain line) shows an EACVF$_{tot}$ value of 0.83 and $d_{1/2}$ value of 0.15 min. The relative underestimation error due to lower EACVF$_{tot}$ value, 0.47/0.83 = 0.57, corresponds precisely to the ratio between the respective $d_{1/2}$ values, 0.15/0.26 = 0.57. Therefore, a correct estimation of n_{max} can be obtained by introducing the real $d_{1/2}$ value into Eq. (10). The present results show that the EACVF$_{tot}$ procedure can take into account, and compensate for, moderate deviation from retention linearity. However, it is clear that a preliminary linearization of the experimental chromatogram is the simplest and most reliable procedure to obtain accurate results.

4.2.2. Signal noise

Another example of unfavorable conditions concerns the presence of baseline noise in the chromatographic signal. Here the simplest case of white noise is considered [35]: noisy chromatographic signals are simulated by adding to the experimental signal white noise with different signal-to-noise ratios (S/N computed by dividing the maximum peak height by three times the noise standard deviation). As an intrinsic property of EACVF$_{tot}$, its value is additive for such independent variables as noise and signal. Therefore, the components of noise and chromatographic signal can be identified in the EACVF$_{tot}$ of the noisy chromatogram: the noisy component can be subtracted to obtain a reliable estimate of the chromatographic parameters. White noise with different S/N ratio values (5, 10, 20, 100) was added to the chromatogram of the standard mixture containing 20 organic compounds (Fig. 1a) to obtain noisy chromatographic signals; as an example, the noisy chromatogram with S/N = 10 is reported in Fig. 4a. The EACVF$_{tot}$ was computed on different noisy chromatograms and their plots are reported in Fig. 4b. The EACVF$_{tot}$ plots show that all the disturbing effects of the noise accumulate at the origin of the EACVF$_{tot}$ yielding a significant increase in EACVF$_{tot}$ for Δt_R values close to 0: the additive contribution of the noise signals with S/N = 5 and 10 is reported in the inset in Fig. 4b, showing the highest effect for the noisiest signal. Therefore, the contribution of baseline noise can be simply eliminated by extrapolating the EACVF$_{tot}$ value to $\Delta t_R = 0$ from EACVF$_{tot}$ values at $\Delta t_R \geq 0.1$ min and by using this extrapolated value to compute m_{tot} value.

Moreover, the EACVF$_{tot}$ values for the deterministic peaks at $\Delta t_R \geq 0.1$ min are completely unaffected by noise and can be used in computations to filter out the noise and obtain a correct estimate of n_{max}. For both the noisy chromatograms with S/N = 5 and 10, the EACVF$_{tot}$ value at $\Delta t_R = 2.5$ min (Fig. 4b, plain and dashed lines) yields a correct estimate of the number of n-alkanes $n_{max} = 6$. The feature of this procedure offers great promise to overcome the drawbacks in processing noisy signals and can also be extended to handling more complex signal noise present in the experimental chromatograms.

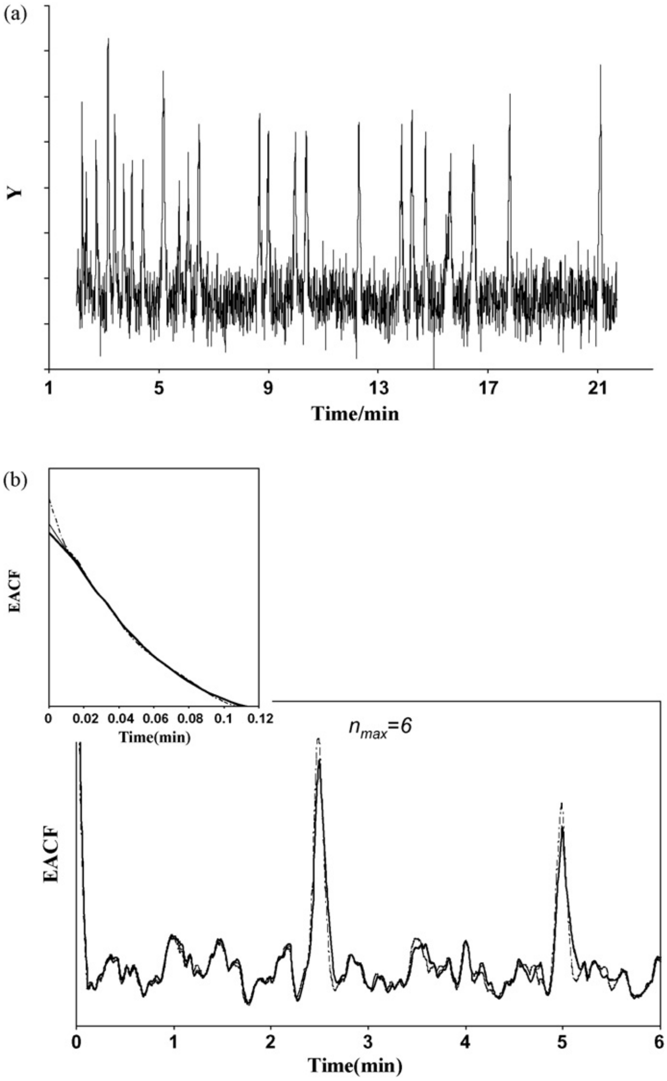


Fig. 4. Noise effect on GC–MS chromatograms of standard mixtures containing 20 organic compounds (Fig. 1b). (a) Noisy chromatogram obtained by superimposition of a white noise with S/N = 10 to the linearized signal (Fig. 1b) and (b) EACVF$_{tot}$ plots computed on the chromatograms. Inset: first region of the EACVF$_{tot}$ for Δt_R values close to 0. Bold line: original chromatogram; dashed line: noisy chromatogram with S/N = 5; plain line: noisy chromatogram with S/N = 10.

4.3. Application to unknown samples

The applicability of the method was tested on a sample containing an unknown number of organics. An ASTM D2887 Reference Oil (b.p. 112–475 °C) was selected as petrochemical sample to represent very complex multi-class mixtures: the total number of compounds has been estimated to exceed one million,

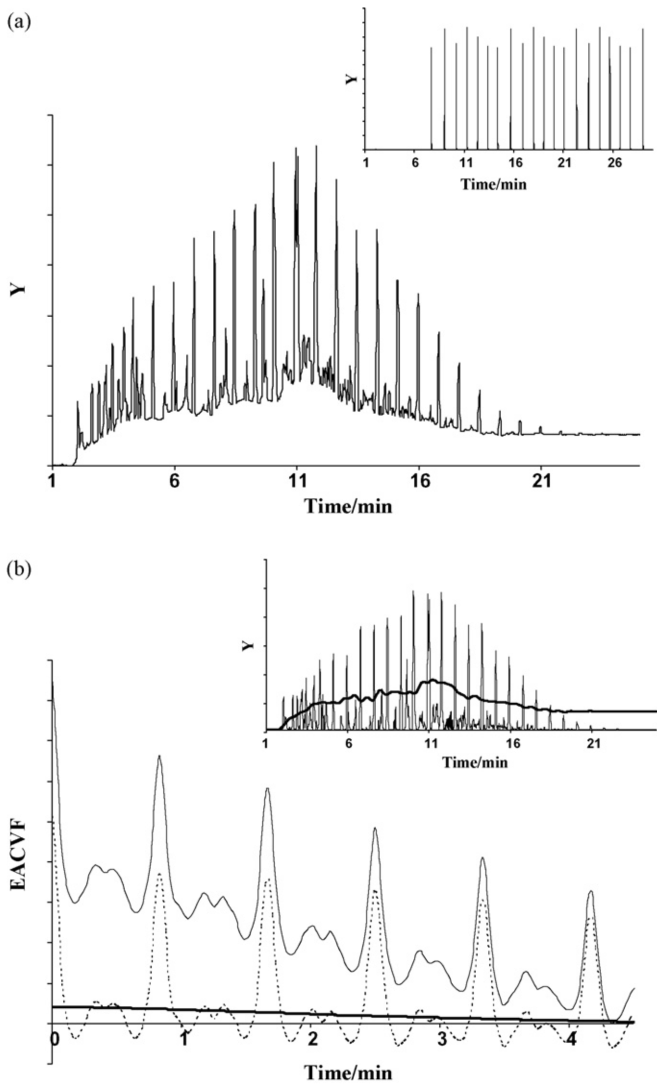


Fig. 5. GC–MS chromatogram of the ASTM D2887 Reference Oil. (a) GC–MS linearized chromatogram. Inset: chromatogram of the reference mixture containing C_6–C_{32} *n*-alkanes. (b) EACVF plots computed on the chromatograms. Plain line: $EACVF_{tot}$ computed on the total signal (Fig. 5a); bold line: $EACVF_{UCM}$ computed on the signal corresponding to UCM; lower dotted line: $EACVF_{res}$ computed on the signal corresponding to the resolved components. Inset: deconvoluted signals corresponding to UCM (bold line) and the resolved components (plain line).

of course many of them are below the detection limits of normal analytical separations [36,37]. The chemical composition characterization of the hydrocarbons, broken down into classes (i.e., paraffinic, olefinic, naphthenic and aromatic), and the amount of hetero-atoms are the basic characteristics that determine the properties of the product [23–25].

The ASTM D2887 Reference Oil was submitted to GC–MS analysis under the temperature-programmed conditions found to be close to the optimal temperature program: 30 °C for 2 min, an increase to 180 °C at $15^\circ\,\mathrm{min}^{-1}$, then to 350 °C at $10^\circ\,\mathrm{min}^{-1}$. The obtained TIC chromatogram (Fig. 5a) clearly shows the typical feature of petrochemical products characterized by the unresolved complex mixture envelope of branched, cyclic and unsaturated hydrocarbons [25,27,28]. A standard mixture containing C_6–C_{32} *n*-alkanes was analyzed under the same chromatographic conditions and the obtained signal (inset in Fig. 5a) served as the basis for the linearization procedure, selecting a value $\Delta t_R = 0.8$ min as constant retention increment between subsequent terms of the series (b in Eq. (3)). The $EACVF_{tot}$ was computed on the linearized signal (Fig. 5b, upper plain line). From the $EACVF_{tot}(0)$ value the SC number present in the mixture was estimated (Eq. (2)): a value $m_{tot} = 115 \pm 10$ was obtained assuming a threshold level of 1% of the highest peak (Table 1, 1st row).

The $EACVF_{tot}$ plot (Fig. 5b, upper plain line) shows well-defined deterministic peaks at $\Delta t_R = 0.8$ min and multiple values that are diagnostic for the presence of homologous series displaying a CH_2 group increment under these experimental conditions. Moreover, the $EACVF_{tot}$ plot shows a specific pattern since it retains information on the UCM band present in the original chromatogram. In order to investigate the chemical composition of the mixture, the baseline corresponding to the UCM signal was subtracted from the total chromatogram (Fig. 5a) with commercial chromatogram processing software [38]: two separated signals corresponding to the UCM (inset in Fig. 5b, bold line) and the resolved components (inset in Fig. 5b, plain line) were obtained. The area of the chromatograms were computed and used to determine the relative abundance of the UCM component versus the total mixture, UCM%: a value of UCM% = 44% was obtained.

On both the signals the EACVF was separately computed to obtain $EACVF_{UCM}$ on the UCM (Fig. 5a, bold line) and $EACVF_{res}$ on the resolved components (Fig. 5b, lower dotted line). The obtained plots clearly show that the $EACVF_{tot}$ computed on the total signal (Fig. 5b, upper plain line) is a combination of $EACVF_{UCM}$ and $EACVF_{res}$. They were then studied to extract quantitative information on the mixture composition (results reported in Table 1, 2nd row). The number of components can be estimated (Eq. (2), using $\sigma^2_{M,tot}/a^2_{M,tot}$ and $d_{h/2}$ from the total chromatogram) from $EACVF_{UCM}(0)$ (bold line in Fig. 5b): a value $m_{UCM} = 35 \pm 6$ is obtained. From $EACVF_{res}(0)$ computed on the resolved components signal (dotted line in Fig. 5b) a value of $m_{res} = 80 \pm 9$ is estimated. The deterministic peaks of $EACVF_{res}$ at $\Delta t_R = 0.8$ min and multiple values, diagnostic for terms of the homologous series, were studied in order to obtain quantitative information on the mixture composition. From $EACVF_{res}$ (0.8 min) the number of terms in the homologous series, $n_{max,ap}$, is estimated as $n_{max,ap} = 56$.

The obtained results (Table 1, 1st and 2nd rows) show that a full characterization of the gas oil chemical composition can be obtained by applying the EACVF method on the original, $EACVF_{tot}$, and on the deconvoluted signals, $EACVF_{UCM}$ and $EACVF_{res}$.

In order to check the results obtained from the EACVF method, the standard procedure, conventionally used to obtain information on the quantitative composition of petrochemicals, was also applied to the experimental chromatogram. This procedure is very laborious and time-consuming since it requires integration of the whole chromatogram and any separated peaks, identification of peaks formed by *n*-alkanes, computation of the total area of the different components, i.e., A_{tot}, A_{res}, A_{UCM} and

Table 1
Comparison of the results obtained from gas oil chromatogram (Fig. 5a) computed with different calculation methods

Method	m_{tot}	m_{UCM}	$m_{\mathrm{UCM}}\%$	m_{res}	$n_{\mathrm{max,ap}}$	n_{max}
$\mathrm{EACVF_{tot}}$	115 ± 10					
$\mathrm{EACVF_{UCM}}$ and $\mathrm{EACVF_{res}}$		35 ± 6		80 ± 9	56	
Conventional			45	86[a]	54	31

[a] Number of peaks counted.

the area of *n*-alkanes. From it $m_{\mathrm{UCM}}\%$ can be computed as the relative area of UCM versus the total area; n_{max} can be counted from the detected peaks identified by MS spectra as terms of the homologous series and $n_{\mathrm{max,ap}}$ can be estimated from the relative area of the terms of the homologous series compared to the total peak area (results reported in Table 1, 3rd row).

The agreement between the results independently obtained with the EACVF method, based on $\mathrm{EACVF_{tot}}$, $\mathrm{EACVF_{UCM}}$ and $\mathrm{EACVF_{res}}$, and the conventional method (compare 1st–3rd rows in Table 1) provides experimental evidence of the reliability of the procedure. One must note the discrepancy between the apparent $n_{\mathrm{max,ap}}$ and the true n_{max} values: this is due to the deviation from reality of the assumption $A_{m,\mathrm{tot}} \approx A_{n_{\mathrm{max}},\mathrm{O}}$ (Eq. (6)) that is the basis to compute $n_{\mathrm{max,ap}}$. The advantage of the EACVF procedure is its ability to extract relevant parameters characterizing a petrochemical sample by computation on the experimental MS-TIC chromatogram.

It must be noted that the obtained parameters concern global information on the chemical composition of the mixture, but the structure elucidation of each peak present in the chromatogram is far from being achieved.

5. Conclusions

The data handling method based on the study of the $\mathrm{EACVF_{tot}}$ has proved to be very useful in classifying compounds into chemical class groups, identifying the presence of homologous series and quantifying the number and abundance of their terms. The introduction of the data pre-processing step to linearize retention time axis increases the method applicability and robustness to investigate complex chromatograms obtained under usual experimental conditions. The whole signal processing procedure makes it possible to achieve a systematic characterization of complex samples by compound class (homologous series, polarity, or functionality): this is the only reliable information that can be drawn from many real-world samples, i.e., natural, industrial or environmental samples where, because of matrix complexity, the separation of all the components is far from being achieved in 1D separations. In such complex samples, given its suitability for group separations, the best technique for analysis of organics is a two-dimensional system, like GC × GC: it is ideal for complex samples containing thousands of compounds but a relatively low number of chemical classes.

If 2D separation equipment is not available, the $\mathrm{EACVF_{tot}}$ data handling method is powerful in extracting relevant information on chemical composition from the resulting, complex 1D chromatogram. In particular, the combination of a proper retention time axis alignment and selective SIM detection magnifies the power of the method in characterizing a much larger fraction of organic compounds.

The above reported results constitute only an example of the wealth of information that can be obtained using the present approach. The method has been further extended to handle the huge amount of data obtained from 2D separations [19,39,40]: study of the 2D-$\mathrm{EACVF_{tot}}$ may form the basis for a comprehensive interpretation of the data matrix acquired in full scan GC–MS analysis containing the whole MS information on component chemical structure.

The method can be proposed to investigate the chemical composition of complex samples of environmental interest: structural class information can be quite useful in monitoring steps in industrial processes and controlling environmental quality, i.e., analyzing biomarkers and environment pollutants in air, water and soil.

Acknowledgements

This work has been supported by the Italian University and Scientific Research Ministry (grant #2005032388_004), and by the University of Ferrara, Italy.

References

[1] E.S. Brodskii, J. Anal. Chem. 57 (2002) 480.
[2] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen, J. Chromatogr. A 1062 (2004) 113.
[3] K. Vekey, J. Chromatogr. A 921 (2001) 227.
[4] J. Krupik, J. Mydlovà, I. Spànik, B. Tienpont, P. Sandra, J. Chromatogr. A 1084 (2005) 80.
[5] P.J. Shoenmakers, J.L.M.M. Oomen, J. Blomberg, W. Genuit, G. van Velzen, J. Chromatogr. A 892 (2000) 29.
[6] C. von Mulhen, C. Alcaraz Zini, E. Bastos Caramao, P.J. Marriott, J. Chromatogr. A 1105 (2006) 39.
[7] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154.
[8] I.G. Zenkevich, Chem. Int. Lab. Syst. 72 (2004) 233.
[9] X. Shao, G. Wang, S. Wang, Q. Su, Anal. Chem. 76 (2004) 5143.
[10] P. Jonsson, J. Guilberg, A. Nordstrom, M. Kusano, M. Kowalczyk, M. Sjostrom, T. Moritz, Anal. Chem. 76 (2004) 1738.
[11] M. Katajamaa, M. Oresic, BMC Bioinformatics 6 (2005) 179.
[12] F. Dondi, A. Betti, L. Pasti, M.C. Pietrogrande, A. Felinger, Anal. Chem. 65 (1993) 2209.
[13] M.C. Pietrogrande, F. Dondi, A. Felinger, J. High Resolut. Chromatogr. 19 (1996) 327.
[14] F. Dondi, M.C. Pietrogrande, A. Felinger, Chromatographia 45 (1997) 435.
[15] A. Felinger, M.C. Pietrogrande, Anal. Chem. 73 (2001) 618A.
[16] M.C. Pietrogrande, P. Coll, R. Sternberg, C. Szopa, R. Navarro-Gonzalez, C. Vidal-Majar, F. Dondi, J. Chromatogr. A 939 (2001) 69.
[17] M.C. Pietrogrande, I. Tellini, A. Fellinger, F. Dondi, C. Szopa, R. Sternberg, C. Vidal-Madjar, F. Dondi, J. Sep. Sci. 26 (2003) 569.
[18] M.C. Pietrogrande, I. Tellini, L. Pasti, F. Dondi, C. Szopa, R. Sternberg, C. Vidal-Madjar, F. Dondi, J. Chromatogr. A 1002 (2003) 179.

[19] N. Marchetti, A. Felinger, L. Pasti, M.C. Pietrogrande, F. Dondi, Anal. Chem. 76 (2004) 3055.
[20] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, Ann. Chim. (Rome) 94 (2004) 721.
[21] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, C. Szopa, R. Sternberg, A. Buch, F. Dondi, J. Chromatogr. A 1071 (2005) 255.
[22] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, Anal. Chem. 78 (2006) 2579.
[23] J. Beens, U.A.Th. Brinkman, Trends in Anal. Chem. 19 (2000) 260.
[24] J. Blomberg, P.J. Shoenmakers, U.A.Th. Brinkman, J. Chromatogr A 972 (2002) 137.
[25] Z. Wang, M.V. Fingas, Mar. Pollut. Bull. 47 (2003) 423.
[26] M.A. Mazurek, Environ. Health Persp. 110 (2002) 995.
[27] F. Liang, M. Lu, T.C. Keener, Z. Liu, S.-J. Khang, J. Environ. Monit. 7 (2005) 983.
[28] A. Cincinelli, S. Mandorlo, R.M. Dickhut, L. Lepri, Atmospheric Environ. 37 (2003) 3125.
[29] K. Kawamura, Y. Imai, L.A. Barrie, Atmospheric Environ. 39 (2005) 599.
[30] K. Kawamura, O. Yasui, Atmospheric Environ. 39 (2005) 1945.
[31] B.L. Karger, L.R. Snyder, C. Horvath, An Introduction to Separation Science, J. Wiley, New York, 1973.
[32] C. Giddings, Unified Separation Science, J. Wiley & Sons. Inc., New York, 1991.
[33] J. Harangi, J. Chromatogr. A 993 (2003) 187.
[34] A. Skvortsov, B. Trathnigg, J. Chromatogr. A 1015 (2003) 31.
[35] A. Felinger, Data Analysis and Signal Processing in Chromatography, Elsevier, Amsterdam, 1998.
[36] D.C. Villalanti, J.C. Raia, J.B. Maynard, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd., Chichester, 2000, pp. 6726–6741
[37] J. Curvers, P. van den Engel, J. High Resolut. Chromatogr. 12 (2005) 16.
[38] PeakFit Software, Aspire Software International, Ashburn, VA.
[39] M.C. Pietrogrande, N. Marchetti, A. Tosi, F. Dondi, P.G. Rigetti, Electrophoresis 26 (2005) 2739.
[40] M.C. Pietrogrande, N. Marchetti, F. Dondi, J. Chromatogr. B 833 (2006) 51.

PAPER II

Data handling of complex GC-MS signals to characterize homologous series as organic source tracers in atmospheric aerosols

Data handling of complex GC-MS signals to characterize homologous series as organic source tracers in atmospheric aerosols

M. C. Pietrogrande, M. Mercuriali, D. Bacco
Department of Chemistry, University of Ferrara, Via L. Borsari, 46, I-44100 Ferrara, Italy. E. mail: mpc@dns.unife.it

Abstract

A description is given of a chemometric approach used to extract information on the characteristics of n-alkane and n-alkanoic acid homologous series as useful markers for PM source identification and differentiation. The key parameters of the homologous series -- number of terms and Carbon Preference Index -- are directly estimated by the Autocovariance Function ($EACVF$) computed on the acquired chromatogram. The homologous series properties — relevant as chemical signature of specific input sources — can be efficiently extracted from the complex CG-MS signal thus reducing the labour and time consumption and the subjectivity introduced by human intervention.

Keywords: aerosol chemical composition/homologous series/GC-MS analysis/ signal processing/ multicomponent mixtures /

1 Introduction

Atmospheric aerosols consist of a complex mixture of hundreds of compounds belonging to many different compound classes: despite this complexity, in environmental monitoring and assessment studies, the sample chemical analysis is usually limited to selected compounds to adequately represent a chemical signature of the possible input sources [1-3]. Homologous series of n-alkanes and n-alcanoic acids are especially suited for use as molecular tracers: they are common to multiple sources and they give information relevant to differentiating aerosols of anthropogenic origin (i.e. associated with industrial and urban activities) from those of natural, biogenic origin [4-6]. The key parameters to characterize specific sources are the number of terms and the carbon preference index (CPI, i.e., the sum of the concentrations of the odd/even carbon number terms divided by the sum of the concentrations of the even/odd carbon number terms). This parameter makes it possible to identify the biogenic contribution (that exhibits a strong odd/even carbon number predominance and thus, a high CPI value) versus petroleum-derived fuels (displaying CPI values close to 1).

Gas chromatography-mass spectrometry (GC-MS), the best analytical technique for these organics, generates extensive amounts of data when applied to such complex mixtures as polluted environmental samples, which are complicated by a vast amount of noise, artefacts, and data redundancy. This motivates the need for computer-assisted signal processing procedures to transform GC data into usable information by extracting all the analytical results hidden in the complex chromatogram [7].

In the present paper, a signal processing procedure based on the AutoCovariance Function ($ACVF$) is applied to GC-MS signals of atmosferic aerosols. The case of n-alkanes and n-alkanoic acids is discussed as useful markers for PM source identification and differentiation. As molecular marker -- number of terms and the CPI value -- the key parameters of the homologous series are directly estimated from the $ACVF$ computed on the acquired chromatogram, thus reducing the labour, time requirements and the subjectivity introduced by human intervention.

1.1 Theory

The chemometric approach studies the Autocovariance Function, $ACVF_{tot}$, that can be directly computed from the experimental chromatogram acquired in digitized form, Experimental $ACVF_{tot}$, $EACVF_{tot}$ [7]. The $EACVF_{tot}$ is plotted vs. the interdistance between subsequent points in the chromatogram Δt to obtain the $EACVF_{tot}$ plot (inset in Fig.1 shows the $EACVF_{tot}$ plot computed on the chromatogram of Fig.1). Theoretical models have been developed to extract information on sample complexity and chromatographic separation from the $EACVF_{tot}$. The mathematical description is reported elsewhere [7-9]: here the main parameters relevant for environmental analysis are discussed:

1. Information on sample complexity and separation performance is contained in the first part of the $EACVF_{tot}$ plot: the number of compounds present in the mixture is estimated from the $EACVF_{tot}$ peak height, and the mean separation performance, σ, from the $EACVF_{tot}$ peak width at half height [7].

2. Information on the separation pattern is contained in the second part of the

$EACVF_{tot}$ plot. In particular, the $EACVF_{tot}$ plot is specifically useful to single out the presence of ordered sequences of peaks appearing in the chromatogram [7]. This is the case of homologous series: if n compounds belonging to a homologous series are present in the sample, they will appear in the chromatogram as an ordered sequence of n peaks located at a constant interdistance value between subsequent terms in the series, e.g., $\Delta t = b$ where b is the CH_2 retention time increment (signed by arrows in the chromatogram of Fig.1) in GC analysis under linearized temperature programming conditions [7]. In this case, the $EACVF_{tot}$ computed on the acquired signal displays well defined deterministic peaks located at the interdistances $\Delta t = bk$, where $k = 1,2,...n-1$ (arrows in the inset of Fig.1): their appearance identifies the presence of the series, even if the corresponding chromatographic peaks are hidden within the complex signal [7].

3. Number of terms of the homologous series. The height of the $EACVF_{tot}$ peaks ($EACVF_{tot}$ values at $\Delta t = bk$) can be quantitatively related to the abundance of the terms of the homologous series, i.e., the combination of the number of terms in the series, n, and their concentration in the sample, according to the following equation:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma a_h^2 (n-k)}{X}\left[\frac{\sigma_h^2}{a_h^2}+1\right] k = 0,1.2.....n-1 \qquad (1)$$

where all the reported parameters can be directly estimated from the chromatographic signal: X is the total chromatogram time span, σ_h^2 / a_h^2 is the peak height dispersion ratio describing the relative abundance distribution of the n terms of the series: it derives from the mean, a_h^2, and the variance, σ_h^2, of peak height computed from the observed peak maxima in the chromatogram [7].

4. Abundance distribution of the homologous series terms A random distribution of the series terms (no odd/even prevalence) yields a monomodal distribution of the subsequent $EACVF_{tot}(bk)$ peaks. If the terms of the series display an odd/even prevalence, the obtained $EACVF_{tot}(bk)$ peaks show a bimodal height distribution with lower values at $\Delta t = bk$ for odd k values and higher values at even k values. This pattern is the basis for extracting quantitative information on the odd/even prevalence of the terms by computing the preference index CPI [9]. Such a parameter can be related to the $EACVF_{tot}(bk)$ values at $\Delta t = b$ and at $\Delta t = 2b$ according to the equation:

$$\frac{EACVF_{tot}(b)}{EACVF_{tot}(2b)} = \frac{\frac{2}{CPI}(n-1)}{\left(1+\frac{1}{CPI^2}\right)(n-2)} \qquad (2)$$

This is a quadratic equation, and can be solved to estimate CPI. The CPI_{tot} value is obtained from $EACVF_{tot}$ by evaluating all the n-alkane components, i.e., the $C_{12}-C_{35}$ range. Otherwise, the CPI index can be calculated on selected n-alkanes in order to describe specific contribution of the n-alkane terms of the sample, i.e., the CPI_{plant} parameter is computed on the heavier $C_{25}-C_{35}$ n-alkanes to describe the contribution of plant waxes. CPI_{plant} is directly estimated from the $EACVF_{plant}$ computed on the partial region of the chromatogram containing the selected terms [9].

All these key parameters, used to characterize the homologous series as source chemical signature, can be directly obtained from the $EACVF_{tot}$ computed on the acquired chromatogram, thus reducing the labour, data handling time and removing the subjective step of peak integration. The big advantages of the present procedure becomes obvious when compared with the traditional procedure which requires identification of the homologous series terms by comparison with retention times and MS spectra of the reference standards, integration of the identified peaks, and computation of CPI from the concentrations of the odd and even carbon numbered terms. It must be underlined that labour and time saving in GC-MS signal processing is especially relevant for environmental analysis requiring high-throughput chemical monitoring.

2 Experimental

The aerosol samples ($PM_{2.5}$ and PM_{10}) were collected daily on quartz-fibre filters in an urban (city centre of Bologna, Italy) and rural sites (San Pietro Capofiume, located on a flat, homogeneous terrain of harvested fields, about 40km north east of Bologna) during Spring 2008.

The PM filters were submitted to the traditional approach of solvent extraction and GC-MS analysis for n-alkane determination (procedure reported in [8]). Then the solution was submitted to the derivatization procedure for n-alkanoic acid analysis: $30\mu l$ of bis(trimethylsilyl) trifluoroacetamide (BSTFA) plus 1% trimethylchlorosilane (TMCS) were added to form trimethylsilyl (TMS) derivatives (reaction at 70 °C for 2h) [7]. The GC-MS system was a Scientific Focus-GC (Thermo-Fisher Scientific Milan, Italy) coupled with PolarisQ Ion Trap Mass Spectrometer (Thermo-Fisher, Scientific, Milan, Italy). The column used was a DB-5 column ($L = 30m$, $I.D. = 0.25mm$, $d_f = 0.25\mu m$=0.25)

(J&W Scientific, Rancho Cordova, CA, USA). Proper temperature program conditions were selected for n-alkanes and n-alkanoic acids to obtain linearized temperature programming conditions, i.e., constant CH_2 retention time increment. The mass spectrometer operated in EI mode (positive ion, $70eV$). Three different samples were analyzed for each PM type: the obtained mean values are reported (Table 1) and discussed below.

3 Results and Discussion

3.1 n-alkane series

The aliphatic hydrocarbons present in the PM samples were identified from the SIM (Selected Ion Monitoring) signal using the typical fragments of these compounds at $m/z = 57 + 71 + 85$ (Figs 1 and 2a for urban and rural samples, respectively). The investigated n-alkanes showed a distribution profile resulting from the contribution of vehicular exhaust and lubricant residues (C_{24} or C_{25} n-alkanes) and inputs of biological sources (C_{27}, C_{29}, and C_{31} terms displaying odd carbon number preference).

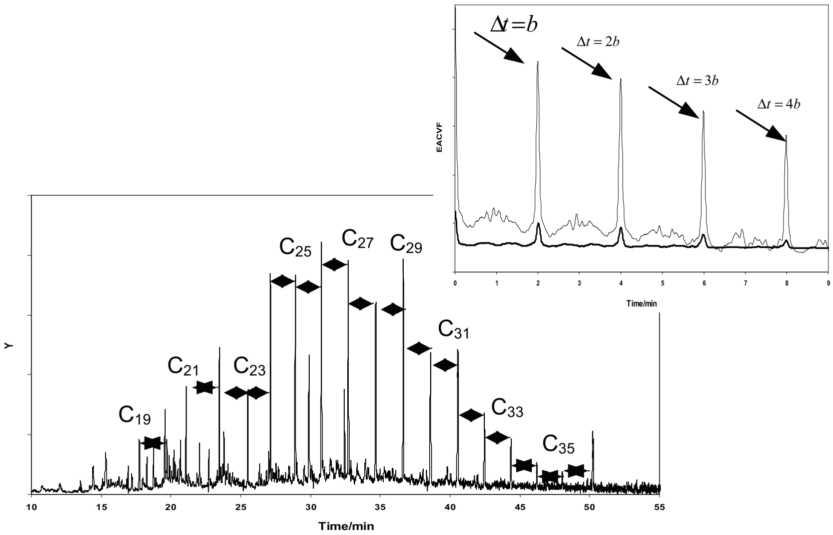


Figure 1. n-alkanes in urban $PM_{2.5}$: GC-MS chromatogram (SIM at $m/z = 57+71+85$); inset: $EACVF_{tot}$ plot (solid line) and $EACVF_{plant}$ plot (bold line).

To extract information on the PM chemical composition, the $EACVF_{tot}$ was computed on the whole chromatographic signal ($EACVF_{tot}$ plots reported in inset of Fig 1 and in Fig. 2b: solid lines). The $EACVF_{tot}$ plots show well-defined deterministic peaks at $\Delta t = 1.9\,\mathrm{min}$ and multiple values that are diagnostic for the presence of the n-alkane homologous series ($b = 1.9\,\mathrm{min}$ in these experimental conditions). The number of n-alkanes present in the mixture, n, can be estimated from the $EACVF_{tot}(1.9\,\mathrm{min})$ values (eq.1): the same value $n = 16$ is obtained from both the chromatograms (Table 1, $EACVF$ estimation).

The $EACVF_{tot}$ values of subsequent peaks give quantitative information on the distribution of the odd/even terms: both the plots show a monomodal distribution of the $EACVF_{tot}$ peak heights suggesting a homogeneous distribution of the odd/even terms. Such a pattern can be quantively described by computing CPI_{tot} (eq. 2): $CPI_{tot} = 1.1$ and $CPI_{tot} = 1.6$ were estimated for urban and rural samples, respectively. These values close to 1 suggest, for both the samples, a major contribution from petroleum residues derived from vehicular emissions as compared to biological inputs.

For all the studied chromatograms, the $EACVF_{tot}$ plots clearly show diagnostic peaks: this behavior highlights the power of the $EACVF$ procedure in extracting information on homologous series, singling them out from the involved signal of the complex chromatograms. In fact, the $EACVF_{tot}$ pattern is independent of the concentration level of n-alkanes, i.e., total concentrations of n-alkanes in the urban $PM_{2.5}$ are nearly four times higher than those in the rural $PM_{2.5}$ sample, and nearly three times lower than those in the urban PM_{10} [4]. Moreover, the chromatographic signal of urban PM samples is further affected by a cluster of unresolved peaks (UCM band) (Fig 1): the $EACVF_{tot}$ of the urban sample (inset in Fig 1, solid line) retains the shape of the UCM band, but clearly displays the $EACVF_{tot}$ peaks characteristic of the homologous series.

To distinguish the role played by the biogenic vs. anthropogenic sources on the atmospheric n-alkanes, the $EACVF_{plant}$ was separately computed on the chromatographic region where the biogenic $C_{27} - C_{35}$ n-alkanes are eluted ($t = 32 - 55\,min$). For both samples, the number of terms $n_{plant} = 9$ is estimated from the $EACVF_{plant}$ values at $\Delta t = 1.9\,min$ ($EACVF_{plant}$ plots in inset of Fig

1 and in Fig 2a, bold lines). The differences in plant contribution to the two samples can be simply identified by visual inspection of the $EACVF_{plant}$ plots obtained. For the rural sample, the $EACVF_{plant}$ (Fig. 2b, bold line) shows a bimodal distribution of subsequent peak heights that is diagnostic for the presence of odd/even prevalence, as revealed by the high estimated value of $CPI_{plant} = 2.4$ that characterizes the contribution of biogenic sources (i.e., higher plant waxes). Otherwise, a lower $CPI_{plant} = 1.3$ value is estimated for the urban sample, as typical for urban environments.

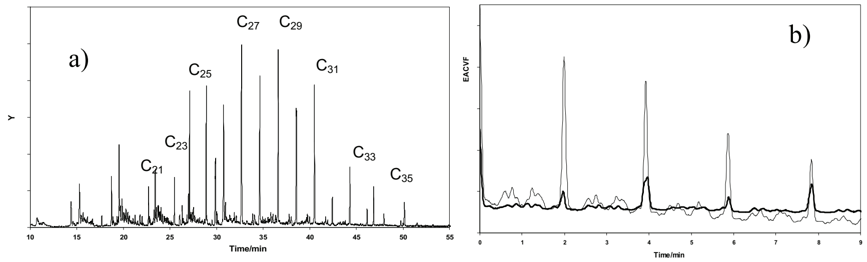


Figure 2. n-alkanes in rural $PM_{2.5}$.
a): GC-MS chromatogram (SIM at $m/z = 57+71+85$);
b): $EACVF_{tot}$ plot (solid line) and $EACVF_{plant}$ plot (bold line).

The $EACVF$ value at $\Delta t = 2b = 3.8\,min$ is related to the total amount of the terms of homologous series (eq.1): therefore, for each sample, the ratio between $EACVF_{tot}(3.8\,min)$ and $EACVF_{plant}(3.8\,min)$ can be used to estimate the relative contribution of plant waxes ($EACVF_{plant}$) to the overall n-alkane components ($EACVF_{tot}$). Such a contribution was quantified as percentages of plant wax fraction in the total n-alkanes: 23% and 10% for the rural and urban samples, respectively.

To check the accuracy of the results obtained (Table 1, 1st-4th columns, $EACVF$ estimation), the traditional procedure, based on computation on the integrated peaks, was applied to the PM chromatograms (Table 1, 5th-8th columns, traditional calculations). A comparison between the independently computed values show a close agreement, validating the reliability of the information obtained by the $EACVF$ procedure. This result confirms the usefulness of the $EACVF$ mehod as a simple, time saving approach to characterize the n-alkane series as molecular biomarker in complex environmental samples.

3.2 n-alkanoic acid series

The $EACVF_{tot}$ method was also applied to characterize n-alkanoic acids, as another homologous series of organics useful in discriminating the relative extent to which various sources contribute to the aerosol burden of organics: the lower molecular weight n-alkanoic acids ($< C_{20}$) are mainly emitted by petroleum based sources, while the heavier $C_{20} - C_{30}$ terms, which display a strong even-to-odd carbon number preference, are mostly derived from plant waxes [6].

After derivatization, the urban and rural PM samples were submitted to GC-MS analysis: the n-alkanoic acids present in the samples were identified in the SIM signal monitoring the typical fragments of the TMS derivatives at $m/z = 75 + 147$ (Fig 3a: rural sample). Under the experimental conditions used, the retention increment for subsequent n-alkanoic acids is $b = 2.5\,min$. The $EACVF_{tot}$ was computed on the whole signal (Fig 3b: solid line): deterministic peaks at $\Delta t = 2.5\,min$ and multiple values are diagnostic for the presence of this homologous series. All the data set to characterize the series are estimated (Table 1, 1st-4th columns, $EACVF$ estimation) and compared to results obtained with the traditional procedure (Table 1, 5th-8th columns, traditional calculations).

Table 1: CPI and n parameters estimated by using $EACVF$ method (1st-4th columns, $EACVF$ estimation) and traditional calculations (5st-8th columns: traditional method).

	$EACVF$ Estimation				traditional method			
Sample	n	CPI_{tot}	n_{plant}	CPI_{plan}	n	CPI_{tot}	n_{plant}	CPI_{plan}
n-alkanes $CPI_{tot} = \sum(C_{13} - C_{35}) / \sum(C_{12} - C_{34})$; $CPI_{plant} = \sum(C_{25} - C_{35}) / \sum(C_{24} - C_{34})$								
PM 2.5 urban	16.2	1.1	8.8	1.3	8.8	1.1	9	2
PM 2.5 rural	15.6	1.6	9.2	2.4	9.2	1.8	9	2.5
PM 10 urban	16.4	0.9	8.6	1.2	8.6	0.9	9	1.8
n-alkanoic acids $CPI_{tot} = \sum(C_{14} - C_{30}) / \sum(C_{13} - C_{29})$; $CPI_{plant} = \sum(C_{20} - C_{30}) / \sum(C_{19} - C_{29})$								
PM 2.5 urban	13.1	6.9	5.6	8.1	14	6.2	6	7.8
PM 2.5 rural	13.6	9.8	7.5	18.7	14	9.4	8	17.2
PM 10 urban	14.2	6.1	5.4	7.9	14	5.4	6	7.5

The $EACVF_{tot}$ plot shows a marked bi-modal distribution with a predominant peak at $\Delta t = 2b = 5\,min$: this is consistent with predominant contribution of hexadecanoic (C_{16}) and octadecanoic (C_{18}) acids that are known to be the most abundant species in most of the PM samples [3, 6]. The even/odd prevalence of acid isomers was confirmed by high $CPI_{tot} = 9.8$ and $CPI_{tot} = 6.9$ values found for rural and urban samples, respectively (Table 1).

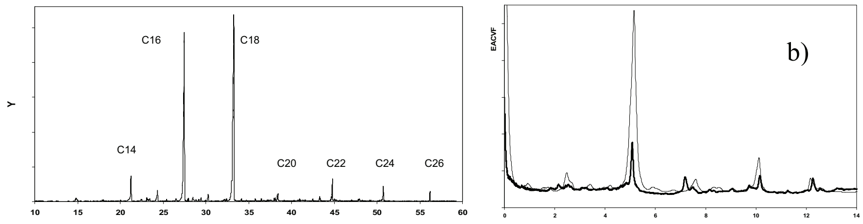


Figure 3: n-alkanoic acids in rural $PM_{2.5}$.
a): GC-MS chromatogram (SIM at $m/z = 75 + 147$);
b): $EACVF_{tot}$ plot (solid line) and - $EACVF_{plant}$ plot (bold line).

To extract information on the biological sources of n-alkanoic acids, the selected chromatographic region containing the $C_{20} - C_{26}$ terms ($35 - 60\,min$) was separately investigated by computing $EACVF_{plant}$. The obtained $EACVF_{plant}$ plot (Fig 3b, bold line) clearly identifies the contribution of biogenic sources, since it displays the strong bi-modal distribution ($EACVF_{plant}(kb)$ peaks are low for odd k and high for even k) characteristic of a strong odd/even prevalence. This is confirmed by the high CPI value ($CPI_{plant} = 18.7$) computed from subsequent $EACVF_{plant}$ peaks, reflecting the stronger vascular plant wax signatures. Otherwise, a lower $CPI_{plant} = 8.1$ value was obtained for the urban PM, indicating that plant waxes make a weaker contribution (Table 1).

The contribution of biogenic n-alkanoic acids in PM samples can also be directly estimated by the ratio between $EACVF_{tot}(5\,min)$ and $EACVF_{plant}(5\,min)$ computed on each chromatogram: the plant fraction ($\geq C_{20}$ congeners) accounted for about 25% and 8% of the total measured n- alkanoic acids levels in rural and urban samples, respectively.

4 Conclusions

The described results reveal the effectiveness of the $EACVF_{tot}$ procedure for handling complex GC-MS data of PM samples in order to characterize the homologous series as molecular marker to trace the origin and fate of atmospheric aerosols. The key parameters -- number of terms and the odd/even prevalence -- are efficiently extracted from the $EACVF$ computed on the acquired chromatogram, with low labor and time consumption. This seems a promising method for high-throughput analysis of the large data sets generated by chemical monitoring in environmental analysis: the obtained chemical information can serve as useful tracers for source apportionment and processes involving organic carbonaceous aerosols when coupled with receptor models.

5 Acknowledgements

The Regional Environmental Agency (ARPAER) is acknowledged for providing the aerosol samples.

6 References

[1] Simoneit, B.R.T., Characterization of organic constituents in aerosols in relation to their origin and transport: a review. International Journal of Environmental Analytical Chemistry 23, 207–237, 1986.

[2] Schauer JJ, Rogge WF, Hildemann LM, Mazurek MA, Cass GR, Simoneit BRT Source apportionment of airborne particulate matter using organic compounds as tracers. Atmospheric Environment 30, 3837–3855, 1996.

[3] Molecular composition of PM2.5 organic aerosol measured at an urban site of Korea during the ACE-Asia campaign Atmospheric Environment 40, 4182–4198, 2006.

[4] Wang G., Liming Huang L., Zhao X., Niu H., Dai Z., Aliphatic and polycyclic aromatic hydrocarbons of atmospheric aerosols in five locations of Nanjing urban area, China, Atmospheric Research, 81, 54– 66, 2006.

[5] Cheng Y., Li S.-M., Leithead A., Brook J.R., Spatial and diurnal distributions of n-alkanes and n-alkan-2-ones on PM2.5 aerosols in the Lower Fraser Valley, Canada Atmospheric Environment, 40, 2706–2720, 2006.

[6] Oliveira C., Pio C:, Alves C., Evtyugina M., Santos P., Goncalves V., Nunes T., Silvestre J.D., Palmgren F., Wahlinc P., Harrad S., Seasonal distribution of polar organic compounds in the urban atmosphere of two large cities from the North and South of Europe, 41, 5555–5570, 2007.

[7] Pietrogrande M.C., Zampolli M.G., Dondi Identification and Quantification of Homologous Series of Compound in Complex Mixtures: Autocovariance Study of GC/MS Chromatograms Analytical Chemistry 78, 2579-2592, 2006.

[8] Pietrogrande M.C., Mercuriali M., Pasti L., Signal processing of GC–MS data of complex environmental samples: Characterization of homologous series Analytica Chimica Acta, 594, 128–138, 2007.

[9] Pietrogrande M.C., Mercuriali M., Pasti L., Dondi F., Data handling of complex GC-MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source identification, submitted to publication.

PAPER III

Data handling of complex GC-MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source

Data handling of complex GC–MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source identification†

Maria Chiara Pietrogrande,* Mattia Mercuriali, Luisa Pasti and Francesco Dondi

Received 3rd September 2008, Accepted 16th December 2008
First published as an Advance Article on the web 14th January 2009
DOI: 10.1039/b815317e

The paper describes a signal method for processing GC–MS signals to extract usable information hidden in the chromatogram thus reducing the labour and time required to handle the data and increasing the quality and objectivity of the results. The method is focused on two relevant parameters for identification and characterization of the n-alkane series present in complex samples (in particular the C_{14}–C_{33} terms): the number of n-alkanes, $n_{\max}$, and the Carbon Preference Index (CPI) describing the odd/even carbon-number predominance. This is a key diagnostic parameter to determine the biogenic and anthropogenic nature of n-alkane sources, useful as chemical markers in source identification and differentiation. The method is a further extension of the approach based on the AutoCovariance Function ($ACVF_{\mathrm{tot}}$): new mathematical equations have been derived and a new computation algorithm implemented to extract information on the n-alkane series – $n_{\max}$ and CPI – directly from the $EACVF_{\mathrm{tot}}$ computed on the acquired chromatographic signal. The procedure was validated on simulated chromatograms where the distribution of the terms of the series describing experimental GC signals was known: the obtained results prove that the parameters $n_{\max}$ and CPI of the homologous series can be estimated with good accuracy and precision. The method applicability was tested on experimental chromatograms of real samples: gasoils and plant extracts were studied to identify n-alkane distribution patterns characteristic of petrogenic and natural samples.

Introduction

Identification and quantification of specific compounds as chemical markers is a convenient approach to characterize the samples formed by a complex mixture of organics. Extensive studies have demonstrated that n-alkanes are especially suited for studies to characterize the origin and fate of different samples; this is because they are widespread components of the environmental carbon cycle and are highly resistant to biochemical degradation and diagenesis in the sedimentary record.[1]

In particular, two parameters are mainly relevant as the chemical signature: (i) the chain length, *i.e.*, average value and maximum carbon number ($C_{\max}$), and (ii) the abundance distribution of the odd/even terms of the series. One common parameter derived from this predominance is the carbon preference index, CPI: it is computed as the ratio of the sum of odd carbon number n-alkanes *vs.* the sum of even carbon number n-alkanes.[2]

The CPI is a key diagnostic parameter to determine the biogenic and anthropogenic nature of sources of n-alkanes: hydrocarbons composed of a mixture of compounds originating from terrestrial plant material show a predominance of odd-numbered carbon chains with $CPI \approx 5$–10,[3,4] whereas petrogenic inputs have a CPI approximating 1.0.[5–8] CPI values close to one are also thought to indicate greater input from marine microorganisms and/or recycled organic matter.[9]

The CPI has proved of great value in environmental and paleo-environmental biomarker-based research in qualitatively and semi-quantitatively apportioning sources of hydrocarbons found in aquatic sediments: the n-alkane distribution pattern is a biomarker which proves helpful in tracking the origins of organic inputs (biogenic or anthropogenic) and identifying 'hot spots' of hydrocarbon contamination.[10]

In petrochemistry, n-alkanes are important constituents of petroleum crudes and their transformation products and thus they are useful tools in oil–oil correlation studies because they provide information regarding an oil, its source rock, genetic associations and alteration.[11] In organic geochemistry, the CPI is used to indicate the degree of diagenesis of straight-chain geolipids, and to numerically represent how much of the original biological chain length specificity is preserved in geological samples.[5,7,10]

Moreover, the chemical characterization of n-alkane constituents of leaf wax coatings has proved to be a quick, reliable and inexpensive method for assessing preliminary chemotaxonomic relationships for systematic classification of plant groups, in combination with other chemical and molecular data: the chemotaxonomic significance of wax alkanes has been demonstrated in studies of many plants groups.[12–16]

Gas chromatography coupled with mass spectrometry (GC–MS) is the well-established technique of choice for identifying and quantifying the hydrocarbon fraction in complex mixtures of

---

Department of Chemistry, University of Ferrara, Via L. Borsari, 46, I-44100 Ferrara, Italy. E-mail: mpc@unife.it

† Electronic supplementary information (ESI) available: Appendix: mathematical derivation of main equations derived to handle the experimental autocovariance function, $EACVF_{\mathrm{tot}}$, computed on the total chromatogram. See DOI: 10.1039/b815317e

organics such as those present in polluted environmental samples (*e.g.* soil, water, aerosol and biota).[17–19] Given the complexity of the samples, complete resolution of all compounds is extremely rare – even when optimum selectivity and extremely high performance separation columns are used; also, GC–MS analyses generate extensive amounts of data. Despite this inherent complexity, only a small percentage of the total number of compounds is usually considered for environmental monitoring and assessment studies: consequently, computer-assisted signal processing is needed to transform GC data into usable information by extracting all the analytical information hidden in the chromatogram, in other words by 'decoding' the complex chromatogram.[20–22] Among the many signal processing procedures developed for this problem, a chemometric approach based on the AutoCovariance Function (ACVF) has been developed and widely applied to experimental chromatograms, a powerful tool for interpreting chromatograms of complex mixtures.[23–27]

Here, this method is further extended: new mathematical equations have been derived and a new algorithm implemented to extract information on the n-alkane distribution pattern. The number of terms and the *CPI* value are directly computed from the ACVF computed on the acquired chromatogram, reducing the labour and time required as well as the subjectivity introduced by human intervention.

The method was validated using PC-generated chromatograms to simulate experimental GC signals of real samples with known odd/even prevalence of the ordered series. The method applicability was tested on real samples representing known anthropogenic and biogenic sources: the results obtained are discussed in terms of the concentration and distribution of n-alkanes as a useful marker for n-alkane source identification and differentiation.

Theory

The chemometric approach studies the AutoCovariance Function ($ACVF_{\mathrm{tot}}$) that can be directly computed from the experimental chromatogram acquired in digitized form. The Experimental $ACVF_{\mathrm{tot}}$ ($EACVF_{\mathrm{tot}}$) at the correlation time Δt is given by the following expression:[24]

$$EACVF_{\mathrm{tot}}(\Delta t) = \frac{1}{N_{\mathrm{p}}} \sum_{j=1}^{N_{\mathrm{p}}-s} \left(Y_j - \bar{Y}\right) \cdot \left(Y_{j+s} - \bar{Y}\right) \qquad s = 0, 1, 2 \ldots M - 1 \tag{1}$$

where Y_j is the digitized chromatogram signal, $\bar{Y}$ its mean value, N_{p} the number of points of the digitized chromatogram, and M the truncation point in the $EACVF_{\mathrm{tot}}$ computation. The correlation time $\Delta t = s\tau$, where τ is the time interval between the subsequent digitized positions, and assumes discrete values with s ranging from 0 to M.

$EACVF_{\mathrm{tot}}$ represents the correlations between subsequent peaks in the chromatogram. Theoretical expressions have been developed to express $ACVF_{\mathrm{tot}}$ in terms of the separation parameters, *i.e.* the number of Single Components (SCs), m_{tot}, the SC peak standard deviation, σ, the distribution of the SC retention pattern (Interdistance Model, IM) and abundance (Abundance Model, AM). Therefore, the $EACVF_{\mathrm{tot}}$ computed on the whole chromatographic signal is the basis for a direct estimation of these parameters, according to the mathematical expressions derived in the previously published studies.[23,28–31]

This study requires theoretical models to describe complex chromatograms: many functions have been developed to describe the infinity of real cases.[28–31] In particular they can be considered as various combinations of the two-limit cases of retention patterns, *i.e.* a Poissonian (P) distribution – a completely disordered separation where SC retention positions are uniform and randomly distributed over the chromatographic axis – and an ordered (O) distribution. SC retention positions are an ordered sequence displaying constant interdistances between subsequent peaks. The simplest approach assumes chromatographic peaks of Gaussian shape with constant width, *i.e.* constant standard deviation σ: this assumption is usually true under optimized programmed temperature conditions.

In the most general case a multicomponent mixture contains m_{tot} SCs with uncorrelated chemical structures that display a Poissonian retention pattern.[23,28] If some of them, n_{max}, belong to a homologous series, they will appear in the chromatogram as an ordered sequence of n_{max} SC peaks where the retention time (t_{R}) of the nth term is described by:

$$t_{\mathrm{R}}(n) = c + bn \qquad n = 0,1,2,3 \ldots n_{\mathrm{max}} \tag{2}$$

where c represents the contribution of a specific functional group to the overall retention, and b is the retention increment between terms of the homologous series, *e.g.* the CH_2 retention time increment, in the strict case of GC analysis under optimized linearized temperature programming conditions.[26] If this condition is not met in practice, a linearization algorithm can be applied to rescale the original signal in order to obtain the same peak width (σ values) and constant retention increment between subsequent terms of the series.[25]

In this case the $EACVF_{\mathrm{tot}}$ method has proved particularly efficient in identifying the presence, and quantify the relative abundance, of the terms of the homologous series. In fact, the $EACVF_{\mathrm{tot}}$ plot displays well-defined deterministic peaks located at interdistances bk, where $b = 1,2,\ldots(n_{\mathrm{max}} - 1)$: their appearance is diagnostic for the presence of the series and their height ($EACVF_{\mathrm{tot}}(bk)$, *i.e.* the $EACVF_{\mathrm{tot}}$ value computed at $\Delta t = bk$), is the basis for estimating the number of SCs belonging to the ordered series, n_{max}, according to the equation:

$$EACVF_{\mathrm{tot}}(bk) = \frac{\sqrt{\pi}\sigma a_{\mathrm{h}}^2 (n_{\mathrm{max}} - k)}{X} \left[\frac{\sigma_{\mathrm{h}}^2}{a_{\mathrm{h}}^2} + 1 \right] \qquad k = 0, 1, 2, 3 \ldots n_{\mathrm{max}} - 1 \tag{3}$$

where σ is the standard deviation of the Gaussian SC peak shape, X the total time range of the chromatogram, a_{h} and σ_{h}^2 are the mean value and the variance of the SC peak heights in the chromatogram to yield the ratio $\sigma_{\mathrm{h}}^2/a_{\mathrm{h}}^2$, called the SC peak height dispersion.[26]

It must be noted that the present procedure makes it possible to identify and characterize the terms of the series only using retention time data, not requiring any information on mass spectra.[24,26]

Abundance distribution of the terms of a homologous series

Here the mathematical model is developed to relate the $EACVF_{\mathrm{tot}}$, computed on the whole chromatogram, to the

This journal is © The Royal Society of Chemistry 2009

abundance distribution of the terms of a homologous series present in the mixture: an algorithm has been developed to estimate the carbon preference index (*CPI*) from $EACVF_{\text{tot}}$.

In order to develop the model, the chromatogram of the homologous series (eqn (2)) is described as the combination of two sequences of peaks representing odd and even terms of the homologous series; they are located at a double repeated interdistance $\Delta t = 2b$ shifted by the quantity b (odd ('o') and even ('e') sequences containing n_{o} and n_{e} peaks, respectively, see Fig. 1a). In this simplified approach the series contains the same number of odd and even terms $n_{\text{o}} = n_{\text{e}} = n$ to yield a total number of terms of the series $n_{\max} = 2n$. The *EACVF* is computed on the signal of each sequence, *i.e.* $EACVF_{\text{o}}$ and $EACVF_{\text{e}}$: their plots show deterministic peaks at the constant interdistance values $\Delta t = 2kb$ according to the following equations (see Fig. 1b):

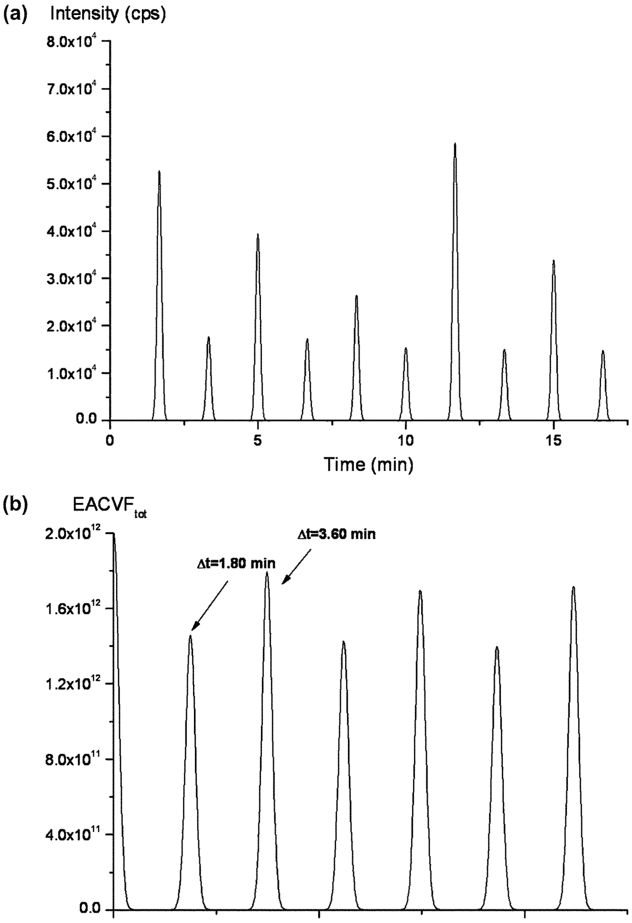


Fig. 1 Simulated chromatogram of a mixture formed by odd and even terms of a sequence ($n_{\text{o}} = n_{\text{e}} = 5$) displaying abundance values generated according to an Exponential AM: mean abundance distributions were simulated for odd and even terms, $a_{\text{o,h}}$ and $a_{\text{e,h}}$, to yield $R = 2$. Each series is formed by 5 terms located at a repeated interdistance $\Delta t = 2b = 3.60$ min shifted by a quantity $\Delta t = b = 1.80$ min. (a): PC-generated chromatographic signal; (b): $EACVF_{\text{tot}}$ plot computed on the signal: the $EACVF_{\text{tot}}$ peaks diagnostic of the sequence at $\Delta t = b = 1.80$ min and $\Delta t = 2b = 3.60$ min are identified by the arrows.

$$EACVF_{\text{o}}(2bk) = \frac{\sqrt{\pi}\sigma a_{\text{o,h}}^2}{X}\left[\frac{\sigma_{\text{o,h}}^2}{a_{\text{o,h}}^2} + 1\right](n_{\text{o}} - k) \qquad k = 1, 2, 3 \ldots n_{\text{o}} - 1 \tag{4}$$

$$EACVF_{\text{e}}(2bk) = \frac{\sqrt{\pi}\sigma a_{\text{e,h}}^2}{X}\left[\frac{\sigma_{\text{e,h}}^2}{a_{\text{e,h}}^2} + 1\right](n_{\text{e}} - k) \qquad k = 1, 2, 3 \ldots n_{\text{e}} - 1 \tag{5}$$

where $a_{\text{o,h}}$, $a_{\text{e,h}}$ and $\sigma_{\text{o,h}}^2$, $\sigma_{\text{e,h}}^2$ are the mean value and the variance of the SC peak heights of the odd and even sequences, respectively.

The whole series containing $n_{\max} = 2n$ terms is obtained by superimposing the 'o' and 'e' sequences and the $EACVF_{\text{tot}}$ computed on the total chromatogram can be investigated as a combination of $EACVF_{\text{o}}$ and $EACVF_{\text{e}}$. New equations are derived to handle the $EACVF_{\text{tot}}$ in order to extract information on the odd/even prevalence of the sequence terms: the main equations are reported in the following and their mathematical derivation is described in detail in the Appendix.†

It is assumed that both the odd and even terms display the same peak abundance distribution, described by peak height dispersion ratio, $\sigma_{\text{h}}^2/a_{\text{h}}^2$, *i.e.*:

$$\frac{\sigma_{\text{o,h}}^2}{a_{\text{o,h}}^2} \approx \frac{\sigma_{\text{e,h}}^2}{a_{\text{e,h}}^2} \approx \frac{\sigma_{\text{h}}^2}{a_{\text{h}}^2} \tag{6}$$

This condition is usually met in real samples since the compound abundances generally follow the most probable Exponential distribution, yielding: $\sigma^2/a^2 \approx 1$. It can be demonstrated that $EACVF_{\text{tot}}(bk)$ values at $\Delta t = bk$ for even k terms are obtained by combining $EACVF_{\text{o}}(bk)$ and $EACVF_{\text{e}}(bk)$ values to yield an equation related to the addition of the two series abundances ($(a_{\text{o,h}}^2 + a_{\text{e,h}}^2)$ term: eqn (a26) in the Appendix†):

$$EACVF_{\text{tot}}(bk) = \frac{\sqrt{\pi}\sigma\left(a_{\text{o,h}}^2 + a_{\text{e,h}}^2\right)(n - k)}{X}\left[\frac{\sigma_{\text{h}}^2}{a_{\text{h}}^2} + 1\right] \qquad k = 0, 2, 4 \ldots 2n - 2 \tag{7}$$

At $\Delta t = bk$ for odd k values, the $EACVF_{\text{tot}}(bk)$ values are given by the cross-correlation term between components of the 'o' and 'e' sequences ($(a_{\text{o,h}} \cdot a_{\text{e,h}})$term: eqn (a28) in the Appendix†):

$$EACVF_{\text{tot}}(bk) = \frac{\sqrt{\pi}\sigma 2(a_{\text{o,h}} \cdot a_{\text{e,h}})(n - k)}{X}\left[\frac{\sigma_{\text{h}}^2}{a_{\text{h}}^2} + 1\right] \qquad k = 1, 3, 5 \ldots 2n - 1 \tag{8}$$

Therefore, the $EACVF_{\text{tot}}(bk)$ peaks computed at subsequent k values give information on the specific abundance distribution pattern of the odd/even terms of the homologous series (Fig. 1b). In fact, if the odd and even terms display the same mean abundance distribution (*i.e.* $a_{\text{o,h}} \approx a_{\text{e,h}}$) eqns (7) and (8) are identical and the $EACVF_{\text{tot}}(bk)$ values are proportional to the values of the sequence $(2n - k)$ for $k = 1,3\ldots(2n - 1)$. Any deviation from such a pattern is diagnostic of the presence of odd/even prevalence among the terms of the series.

For the sake of simplicity, the first $EACVF_{\text{tot}}(bk)$ peaks for $k = 1$ and $k = 2$ are considered below.

To describe a specific odd/even distribution pattern for the terms of the homologous series, the R value is defined as the ratio between the mean value of the SC peak height of odd *vs.* even terms:

$$R = \frac{a_{\mathrm{o,h}}}{a_{\mathrm{e,h}}} \tag{9}$$

By computing eqns (7) and (8) for $k = 1$ and $k = 2$, respectively, and introducing the R parameter, the following equations are obtained:

$$EACVF_{\mathrm{tot}}(2b) = \frac{\sqrt{\pi}\sigma}{X}\left(a_{\mathrm{o,h}}^2\right)\left(1+\frac{1}{R^2}\right)\left[\frac{\sigma_{\mathrm{h}}^2}{a_{\mathrm{h}}^2}+1\right](n-2) \tag{10}$$

$$EACVF_{\mathrm{tot}}(b) = \frac{\sqrt{\pi}\sigma}{X}2a_{\mathrm{o,h}}^2\frac{1}{R}\left[\frac{\sigma_{\mathrm{h}}^2}{a_{\mathrm{h}}^2}+1\right](n-1) \tag{11}$$

By dividing eqn (11) by eqn (10), the following expression can be obtained as a function of R:

$$\frac{EACVF_{\mathrm{tot}}(b)}{EACVF_{\mathrm{tot}}(2b)} = \frac{\frac{2}{R}(n_{\max}-1)}{\left(1+\frac{1}{R^2}\right)(n_{\max}-2)} = \frac{2R(n-1)}{\left(R^2+1\right)(n-2)} \tag{12}$$

The equation can be simplified by introducing the approximation that the ratio between $(n-1)$ and $(n-2)$ is equal to 1: this is strictly true for large n values, *i.e.* $n \to \infty$, otherwise it can be applied once n is known. With this assumption eqn (12) can be simplified into:

$$\frac{EACVF_{\mathrm{tot}}(b)}{EACVF_{\mathrm{tot}}(2b)} = \frac{2R}{\left(R^2+1\right)} \tag{13}$$

This is a simple quadratic equation, that can be solved to obtain the R value:

$$R = \frac{2 \pm \sqrt{4-4Y}}{2Y} \tag{14}$$

where

$$Y = \frac{EACVF_{\mathrm{tot}}(b)}{EACVF_{\mathrm{tot}}(2b)} \tag{15}$$

Eqn (13) shows that the odd/even prevalence of the terms of the homologous series, expressed by the R value, can be directly estimated from the whole chromatogram by computing the $EACVF_{\mathrm{tot}}$ values at $\Delta t = b$ and $\Delta t = 2b$ on the total signal.

Number of the terms of a homologous series

Applying the general model based on eqn (3) to estimate $n_{\max}$ may yield misleading results in the case of an odd/even prevalence: in fact, at $\Delta t = bk$ for odd k values, the $EACVF_{\mathrm{tot}}$ values strongly depend on the presence of the odd/even prevalence in the peak abundance since it is related to the product $(a_{\mathrm{o,h}} \cdot a_{\mathrm{e,h}})$ (eqn (8)). Otherwise, at $\Delta t = bk$ for even k values, the $EACVF_{\mathrm{tot}}$ values are independent of the peak abundance distribution of the odd and even terms since it is related to the quantity $(a_{\mathrm{o,h}}^2 + a_{\mathrm{e,h}}^2)$ (eqn (7)): therefore, at $\Delta t = bk$ for even k values, $EACVF_{\mathrm{tot}}$ values are used to obtain a correct estimation of $n_{\max}$. In order to make the procedure more robust, for even k values, the computation is based on two subsequent $EACVF_{\mathrm{tot}}$ deterministic peaks at $\Delta t = bk$ and $\Delta t = b(k + 2)$ according to the following equation:

$$n_{\max} = 2\frac{EACVF_{\mathrm{tot}}(bk)}{EACVF_{\mathrm{tot}}\left(b(k+2)\right)} + k \tag{16}$$

The mathematical derivation of eqn (16) is reported in detail in the Appendix.† The correct estimation of $n_{\max}$ based on eqn (16) makes it possible to achieve an accurate estimation of R by using the rigorous eqn (12) to remove the approximation introduced in eqn (13). It must be underlined that the mathematic model developed on the basis of eqns (12) and (16), strictly derived for a chromatogram that only contains homologous series terms, is applicable to the general case of complex mixtures containing random uncorrelated compounds in addition to the homologous series. In fact, the Poissonian component yields $EACVF_{\mathrm{tot}}$ values significantly different from 0 only for $\Delta t \leq 4\sigma$, so that, at the repeated interdistances ($\Delta t = bk$), the $EACVF_{\mathrm{tot}}$ values are mainly due to the contribution of the homologous series (eqn (3)) and can be used to evaluate its properties.[27]

Computation of the *CPI* value

The *CPI* can be calculated by using the different n-alkane terms present in the mixture to describe the different nature of the n-alkane component of the sample:[2,17] the whole range of n-alkanes is used to describe the whole n-alkane component:

$$CPI_{\mathrm{tot}} = \frac{\sum(\mathrm{C}_{13}-\mathrm{C}_{35})}{\sum(\mathrm{C}_{12}-\mathrm{C}_{34})} \tag{17}$$

the C_{13}–C_{25} n-alkanes are the basis for describing the petrogenic fraction:

$$CPI_{\mathrm{pet}} = \frac{\sum(\mathrm{C}_{13}-\mathrm{C}_{25})}{\sum(\mathrm{C}_{12}-\mathrm{C}_{24})} \tag{18}$$

the heavier C_{25}–C_{35} n-alkanes are used to describe the biogenic contribution:

$$CPI_{\mathrm{bio}} = \frac{\sum(\mathrm{C}_{25}-\mathrm{C}_{35})}{\sum(\mathrm{C}_{24}-\mathrm{C}_{34})} \tag{19}$$

It must be noted that all the computation procedures are based on the same number of odd and even terms of the n-alkanes, *i.e.* $n_{\mathrm{o}} = n_{\mathrm{e}}$. Therefore, the R value, which is based on the mean peak height of odd *vs.* even terms (eqn (9)), can be properly used to estimate *CPI*: the contribution of selected n-alkanes can be identified by computing the $EACVF_{\mathrm{tot}}$ over a partial region of the chromatogram which has been correctly chosen so that it contains a specific range of n-alkanes.

Experimental

Material and methods

Chemicals and supplies. The standard mixtures of C_{12}–C_{34} n-hydrocarbons were purchased from Supelco (Milan, Italy) (99% min).

The petrogenic samples studied were a commercial diesel fuel containing C_{10}–C_{29} n-alkanes and an ASTM D2887 Reference Oil (Supelco, Milan, Italy). This sample is the basis for the standard test method for the boiling range distribution of petroleum fractions by GC (C_6–C_{44}, bp 115–475 °C).

The natural samples investigated were leaves and flowers derived from three different plants. One was formed by the florist's 'Mimosa': racemose inflorescences made up of numerous smaller, bright yellow globose flowerheads produced by *Acacia dealbata* (Silver Wattle).[32] Another sample was chrysanthemum flowers (*Chrysanthemum coronarium*): a genus of a perennial flowering plant. The third sample was formed by the leaves and flowers of *Hypericum perforatum*. This is a shrub that grows in the wild in temperate regions and is widely used in traditional and official medicine because of its antidepressant, antiviral, and antimicrobial activity.[33,34]

Analytical procedure. The studied oil samples were a commercial diesel fuel and an ASTM D2887 Reference Oil: the samples were properly diluted in iso-octane prior to injection into the GC–MS system.

The studied plant samples were: Mimosa dry flowers, chrysanthemum fresh flowers and *H. perforatum* fresh leaves and flowers. 500 mg of material were extracted twice with 5 mL of dichloromethane (Sigma-Aldrich, Milan, Italy) using ultrasonic agitation for 20 minutes. The extracts were combined, filtered with a PTFE filter (0.45 μm) and then evaporated to dryness by a gentle stream of N_2. The sample was then dissolved in iso-octane (50 μL) and injected into the GC–MS system. A total of three extractions and injections was performed for each sample.

The GC–MS system was a Scientific Focus-GC (Thermo-Fisher Scientific Milan, Italy) coupled with PolarisQ Ion Trap Mass Spectrometer (Thermo-Fisher, Scientific, Milan, Italy). The column used was a DB-5 column ($L = 30$ m, I.D.=0.25 mm, $d_f = 0.25$ μm film thickness) (J&W Scientific, Rancho Cordova, CA, USA). High purity helium was the carrier gas with a velocity of 1.0 mL/min. The temperature program for n-alkane analysis was set as follows: the initial temperature (35 °C) was raised to 120 °C at 7 °C/min, then it was increased to 240 °C at 5 °C/min, then further raised to 320 °C at 3 °C/min. All samples were injected in splitless mode; the injector temperature was 300 °C.

The mass spectrometer operated in EI mode (positive ion, 70 eV): mass spectra were acquired with repetitive scanning from 40 to 400 *m*/*z* in 1 s. Ion source and transfer-line temperatures were 250 °C and 320 °C, respectively. In addition to TIC chromatograms the SIM signals were also monitored by selecting *m*/*z* values of 57, 71 and 85 which are characteristic ion fragments for n-alkanes.

All the n-alkanes were identified by comparison with retention times and mass spectra of reference n-alkane standards (C_{10}–C_{34}).

Computation

The algorithms used for the calculation and for the signal processing of GC–MS data are written in Fortran77 and *MATLAB* and run on a 1.53 GHz (256 RAM), AMD *Athlon* personal computer.

Chromatograms simulation. Chromatograms were generated with a *MATLAB* routine, written in-house to simulate noise-free GC–MS signals. The simulated chromatograms were generated by setting four vector values (*time*, *peakint*, *peaksig* and *peaktr*) where *time* is the time axis vector, *peakint* a vector with peak height, *peaksig* is peak standard deviation vector and *peaktr* the retention time vector.

In order to generate a GC signal of homologous series with known odd-to-even predominance, each chromatogram was computed as a combination of two ordered sequences of deterministic peaks representing the odd and the even components of a homologous series. They are located along the retention time axis at constant interdistance ($\Delta t = 2b = 3.60$ min) and shifted by a quantity ($\Delta t = b = 1.80$ min).

Peak shape was described by a Gaussian function (12 points per peak): a σ value of 0.03 min was assumed as a constant width for all the chromatographic peaks to provide a good description of the experimentally obtained GC signals. Peak height values were generated according to an Exponential AM, where SC abundances are randomly distributed around the mean value $\bar{a}_h$. Moreover, peaks are sorted according to their increasing or decreasing height to obtain a 'self-structured' distribution resembling the peak height pattern commonly present in real samples.[18,35]

The peak height values for the odd/even sequences were properly generated in order to obtain different mean values $a_{o,h}$ and $a_{e,h}$, thus yielding different ratio values (eqn (9)): the *R*-instigated values were 2, 3, 4, and 5.

The reliability of the procedure was verified on two types of simulated chromatograms formed by 50 SCs (each sequence contains 25 terms) or 10 SCs (each sequence contains 5 terms). The latter case closely represents the experimental chromatograms of real samples, where the hydrocarbon fraction is dominated by n-alkanes ranging from C_{13} to C_{34}.[17–19] For each *R* value and peak height distribution, 25 simulated chromatograms were generated with the same *R* and n_{max} parameters: computations were performed on them to evaluate the accuracy and precision of the mathematical procedure, expressed as relative error (ε%) and variation coefficient (*CV*%) of the obtained results.

***EACVF*$_{tot}$ calculation.** The first step in data handling consists of linearizing the chromatographic signal to obtain constant retention increments between subsequent terms of the homologous series (eqn (2)). It is a retention time alignment procedure based on comparison *vs.* a set of n-alkane standards.[25,31] The AutoCovariance Function was then numerically calculated from the linearized chromatogram, according to eqn (1). A *MATLAB* algorithm based on eqns (12) and (16) was implemented to directly estimate the two parameters n_{max} and *R* from the $EACVF_{tot}$ computed on the whole chromatographic signal: by a proper selection of different regions of the chromatogram the computed *R* value corresponds to parameters CPI_{tot}, CPI_{pet} and CPI_{bio}.

Results and discussion

The proposed method's robustness and reliability in estimating the n_{max} and *R* parameters was verified on simulated chromatograms with a known distribution of the sequence terms. Different peak height distribution models were assumed to generate chromatographic signals describing experimental chromatograms of real samples. The applicability of the method was tested on experimental chromatograms of real samples of anthropogenic and natural origin. The parameters obtained are useful molecular markers for comparing known sources and observed atmospheric samples to identify sources of organic matter emissions.

Validation on PC-generated chromatograms

PC signals were generated to describe chromatograms of mixtures containing homologous series: they were computed by the combination of two sequences of peaks representing odd and even terms of the homologous series. For each series, peak height values were generated according to an Exponential AM, where SC abundances are randomly distributed in a given interval (*phr*) around the mean value a_h: it has been theoretically demonstrated and experimentally verified that this random function is the most likely distribution for the maximum complexity of the mixture, since it contains the maximum entropy. Peak height values are simulated for the two separated sequences in order to obtain known odd/even preference values described by the *CPI* values drawn from 2 to 5.

Chromatogram with a random peak height distribution. Different chromatograms were simulated containing a higher (50) and lower (10) number of homologous series terms displaying known *CPI* values.

As an example, Fig. 1a reports a simulated chromatogram containing 10 terms of a homologous series with a CH_2 retention increment $b = 1.80$ min: the series is obtained by superimposing two separated sequences containing 10 SCs, *i.e.* $n_o = n_e = 5$ with an abundance distribution to yield an odd/even preference $CPI = 2$. The $EACVF_{tot}$ plot computed on the chromatogram (Fig. 1b) clearly shows a specific pattern of subsequent $EACVF_{tot}$ peaks related to the odd/even preference. It consists of two decreasing sequences of peaks: a sequence of lower $EACVF_{tot}(bk)$ peaks at the odd b values containing information about the cross-correlation between odd/even terms (eqn (8)) and a series of higher $EACVF_{tot}(bk)$ peaks at even b values related to the addition of the terms of both the series (eqn (7)).

The implemented algorithm computes the R value directly by comparing $EACVF(1.80$ min) and $EACVF(3.60$ min) values, according to eqn (14), and computes the number of terms in the series, n_{max}, by comparing $EACVF(3.60$ min) and $EACVF(7.20$ min) values, according to eqn (16).

Computations were performed on different PC-generated chromatograms by varying n_{max} (50 and 10), R values (2, 3, 4, 5) and the dispersion of peak abundance distribution (*phr*) (Table 1, theoretical values, 1st–11th rows). For each parameter set, 25 simulated chromatograms were independently generated and computations were performed to estimate R and n_{max} values according to eqns (14) and (16). The mean estimated values and their confidence intervals (at 95% of probability) are reported in Table 1 (calculated values, 1st–11th rows). The accuracy of the computation procedure was estimated as the relative error (ε% values, 5th and 8th columns) and precision, expressed as *CV*% (*CV*%, 6th and 9th columns). The good precision and accuracy of the obtained results may validate the suitability of the developed method to describe the distribution of homologous series terms.

Table 1 Theoretical and $EACVF_{tot}$ calculated parameters on PC-generated chromatographic signals[a]

Random distribution

	Theoretical		Calculated					
phr	R	n_{max}	R	ε%	*CV*%	n_{max}	ε%	*CV*%
0.66/1.33	2	50	2.1 ± 0.16	6.5	3.75	49.0 ± 0.78	2.0	0.81
0.61/1.38	2	50	2.2 ± 0.25	10.0	5.90	49 ± 1.1	3.0	1.19
0.55/1.50	2	50	2.3 ± 0.21	15.0	4.78	47 ± 1.9	6.2	2.02
0.37/1.62	2	50	2.4 ± 0.27	20.0	5.83	45 ± 2.3	10.0	2.66
0.66/1.33	3	50	3.2 ± 0.29	6.6	4.68	49 ± 1.1	2.0	1.16
0.66/1.33	4	50	4.2 ± 0.37	5.0	4.52	49 ± 1.2	2.0	1.26
0.66/1.33	5	50	5.3 ± 0.53	6.0	5.09	49 ± 1.3	2.0	1.35
0.66/1.33	2	10	2.1 ± 0.80	5.0	17.14	10.4 ± 0.65	4.0	2.79
0.66/1.33	3	10	3 ± 1.0	16.6	13.14	10 ± 1.0	1.0	4.45
0.66/1.33	4	10	4 ± 1.1	5.0	10.00	9.9 ± 0.80	1.0	3.64
0.66/1.33	5	10	5 ± 1.7	0.0	15.20	10.2 ± 0.65	2.0	2.84

'Self-structured' distribution

	Theoretical		Calculated					
phr	R	n_{max}	R	ε%	*CV*%	n_{max}	ε%	*CV*%
0.66/1.33	2	50	2.0 ± 0.27	0.0	7.00	48.6 ± 0.94	1.4	0.99
0.61/1.38	2	50	1.9 ± 0.21	5.0	5.79	48 ± 1.1	4.6	1.17
0.55/1.50	2	50	2.0 ± 0.35	0.0	9.00	45 ± 2.0	10.0	2.22
0.37/1.62	2	50	1.9 ± 0.31	5.0	8.42	42 ± 2.1	16.0	2.62
0.66/1.33	3	50	3.0 ± 0.35	0.0	6.00	48.6 ± 0.80	2.8	0.84
0.66/1.33	4	50	4.0 ± 0.33	0.0	4.25	49 ± 1.1	2.8	1.19
0.66/1.33	5	50	5.0 ± 0.52	0.0	5.40	49 ± 1.2	2.4	1.25
0.66/1.33	2	20	1.9 ± 0.50	5.0	12.63	20.7 ± 0.23	3.5	0.53
0.66/1.33	3	20	3.0 ± 0.54	0.0	8.66	20.8 ± 0.33	4.0	0.77
0.66/1.33	4	20	4 ± 1.0	2.5	12.19	20.8 ± 0.27	4.0	0.62
0.66/1.33	5	20	5 ± 1.2	6.0	10.56	20.8 ± 0.31	4.0	0.72
0.66/1.33	2	10	2 ± 1.33	5.0	28.57	9.9 ± 0.89	1.0	4.04
0.66/1.33	3	10	3 ± 1.7	10.0	23.63	9.8 ± 0.93	2.0	4.28
0.66/1.33	4	10	4 ± 1.4	0.0	16.25	10 ± 1.4	1.0	6.56
0.66/1.33	5	10	5 ± 1.8	6.0	15.09	10.0 ± 0.82	0.0	3.70

[a] The confidence intervals at 95% of probability are reported for the mean R and n_{max} estimated values (4th and 7th columns).

Chromatogram with a 'self-structured' peak height distribution. The GC–MS signals of the aliphatic hydrocarbon fraction of environmental samples are usually dominated by n-alkane peaks corresponding to terms ranging from C_{13} to C_{35} with abundance normally distributed around a maximum of the most abundant C_{max} term, thus generating a bell-shaped pattern. Chromatographic signals were generated to specifically resemble such a peak height pattern: in the following, this model is called 'self-structured' distribution. For each homologous sequence, formed by $n_o = n_e$ peaks, the peak height values were generated according to a random distribution and then sorted according to increasing and decreasing heights to yield a bell-shaped distribution centred around the C_{max} value. As an example, Fig. 2a reports a simulated chromatogram containing 20 terms of a homologous series ($b = 1.80$ min) displaying a 'self-structured' distribution and an odd/even preference $R = 2$. The $EACVF_{tot}$ plot computed on the chromatogram (Fig. 2b) clearly shows the recursive pattern diagnostic of the presence of the odd/even preference. The developed algorithm directly computes the R value by comparing $EACVF_{tot}(1.80$ min) and $EACVF_{tot}(3.60$ min) values, according to eqn (14). The $EACVF_{tot}$ peak heights

This journal is © The Royal Society of Chemistry 2009

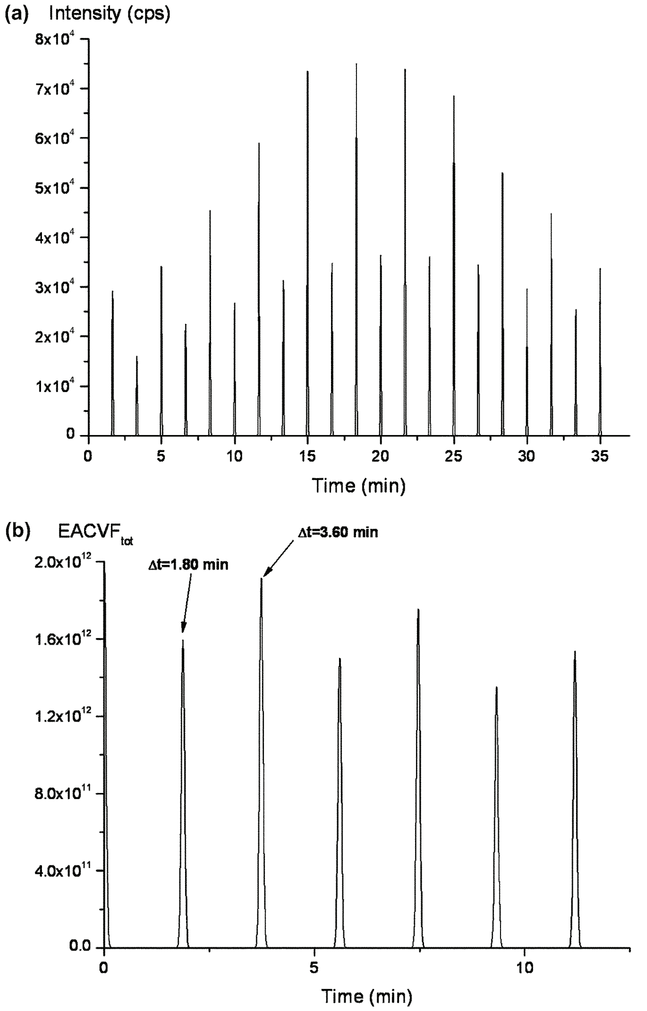


Fig. 2 Simulated chromatogram of a mixture formed by odd and even terms of a sequence ($n_o = n_e = 10$) displaying abundance values generated according to a 'self-structured' distribution resembling peak height pattern commonly present in real samples: mean abundance distributions were generated for odd and even terms, $a_{o,h}$ and $a_{e,h}$, to yield $R = 2$. Each series is formed by 10 terms located at a repeated interdistance $\Delta t = 2b =$ 3.60 min shifted by a quantity $\Delta t = b = 1.80$ min (a): PC-generated chromatographic signal; (b): $EACVF_{tot}$ plot computed on the signal: the arrows identify the peaks diagnostic of the sequence at $\Delta t = b = 1.80$ min and $\Delta t = 2b = 3.60$ min.

depend on the position of the peak along the time axis, displaying a Gaussian pattern following the bell-shaped distribution of the original SC peaks. In this case, a critical point for the correct estimation of n_{max} (eqn (16)) is the selection of the odd k values on which $EACVF_{tot}(bk)$ is computed. To overcome this problem, the computation was performed on different pairs of subsequent peaks, *i.e.* $EACVF_{tot}$(3.60 min) and $EACVF_{tot}$(7.20 min); $EACVF_{tot}$(7.20 min) and $EACVF_{tot}$(10.80 min), and the n_{max} value was estimated as the mean of the results obtained (Table 1, self-structured distribution, 7th column).

Computations were performed on different 'self-structured' chromatograms containing 10, 20 and 50 SCs with different R values (2, 3, 4, 5) and different random peak height distribution (*phr*) (theoretical values in Table 1, 12th–26th rows). Each estimated result for R and n_{max} is the mean of the computation repeated on 25 independently simulated chromatograms with the same parameter set (calculated values in Table 1, 12th–26th rows). A good agreement between the theoretical and calculated values is obtained: the accuracy and precision of the computation procedure were evaluated as relative error, ε%, and variance coefficient, *CV*% (Table 1, 5th, 8th and 6th, 9th columns, respectively). The results obtained validate the reliability of the computation procedure in characterizing the properties of a homologous series distribution that very closely resembles the experimental pattern.

Applications to real samples

The applicability of the method was tested on real samples of known anthropogenic and natural origin such as oil samples and plant extracts. The results obtained are discussed in terms of their relevance as molecular markers for the characterization of possible sources of organic inputs.

Fuel sample. Two oil samples were chosen to test the reliability of the $EACVF_{tot}$ method in investigating the properties of petrogenic or anthropogenic n-alkanes. A *CPI* value close to 1 is a diagnostic indicator of petrogenic hydrocarbon contamination in marine sediments[8] or fossil fuel combustion as primary emission sources for the urban particulate matter.[36]

As an example, the GC–MS signal of the volatile components of a commercial diesel fuel is reported in Fig. 3a: the SIM signal for monitoring the n-alkanes at *m/z* values of 57, 71 and 85 is reported. The n-alkanes were identified using the GC retention times of the reference standards (C_{10}–C_{30}): the main components are mid-chain n-alkanes C_{10}–C_{25}, C_{17} and C_{19} being dominant. A visual examination of the chromatogram shows a typical chromatographic profile of petrogenic n-alkanes characterized by no odd-to-even predominance. The $EACVF_{tot}$ was computed on the whole chromatogram (lower solid line in Fig. 3c): its plot clearly shows a monomodal distribution of the $EACVF_{tot}$ peak height suggesting a homogeneous distribution of the odd/even terms. Such a pattern can be quantified by computing CPI_{tot} according to eqn (14): by selecting the proper retention region containing C_{13}–C_{25} n-alkane ranges, CPI_{pet} can be estimated to characterize the petrogenic fraction present in the sample: CPI_{tot} and CPI_{pet} values close to 1 were obtained (estimated values, 2nd and 3rd columns in Table 2). With the developed algorithm the n_{max} n-alkanes present in the sample can be directly estimated from the $EACVF_{tot}$ peaks at $\Delta t = bk$, even k (eqn (16)) (estimated values, 5th column in Table 2). The data obtained for ASTM D2887 Reference Oil show a similar chemical composition of this sample (3rd row in Table 2).

The accuracy of the results was checked by comparing them with results obtained using the traditional procedure. It requires identification of the n-alkanes by comparison to reference standards and MS spectra, integration of the identified peaks, computation of *CPI* as a ratio of the sum of concentrations of the odd-numbered carbon alkanes *vs.* that of the even-numbered terms. The obtained results (traditional calculations, 6th and 7th columns in Table 2) show a close similarity with data estimated by $EACVF_{tot}$: this agreement is a proof of the usefulness of the procedure for a simple and quick characterization of the n-alkane distribution pattern as a molecular biomarker in complex samples.

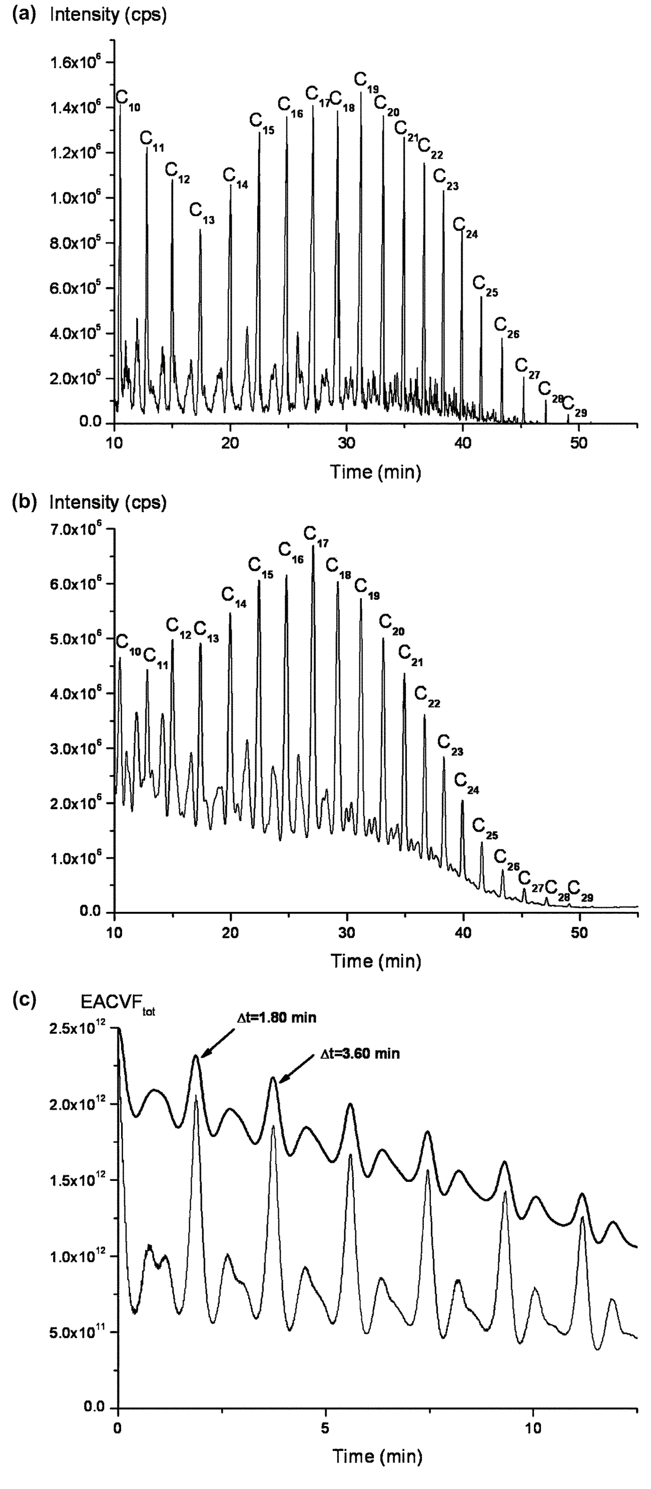


Fig. 3 Volatile components of a commercial diesel fuel (a): SIM signal of the GC–MS chromatogram monitored at $m/z = 57 + 71 + 85$; (b): TIC signal of the GC–MS chromatogram; (c): $EACVF_{\mathrm{tot}}$ plot computed on the whole chromatogram: lower solid line: $EACVF_{\mathrm{tot}}$ on the SIM signal (chromatogram in Fig. 3a); upper bold line: $EACVF_{\mathrm{tot}}$ on the TIC signal (chromatogram in Fig. 3b). The arrows identify the peaks diagnostic of the sequence at $\Delta t = b = 1.80$ min and $\Delta t = 2b = 3.60$ min.

The ability of the $EACVF_{\mathrm{tot}}$ procedure to handle complex signals can be emphasized by extending the investigation to involved TIC signals. The TIC chromatogram of the oil sample was studied (Fig. 3b): it displays the typical chromatographic profile characterized by the UCM band (Unresolved Component Mixture) formed by a cluster of unresolved peaks. The $EACVF_{\mathrm{tot}}$ plot (Fig. 3c, upper bold line) is strongly affected by the specific pattern of the UCM band which is superimposed on the deterministic $EACVF_{\mathrm{tot}}$ peaks, displaying monomodal height distribution. Nevertheless, the $EACVF_{\mathrm{tot}}$ model makes it possible to single out the n-alkane sequence properties by computing the n_{max} and CPI_{tot} values using eqns (16) and (14) on $EACVF_{\mathrm{tot}}$ computed over the whole original signal. The obtained results (2nd row in Table 2) show a close similarity to the data obtained from the SIM signal (1st row in Table 2) and from the traditional calculation method (traditional calculation, 1st and 2nd rows in Table 2). This result confirms the robustness of the developed method in extracting reliable information from the direct handling of complex chromatograms, such as SIM and TIC GC–MS involved signals.

Plant samples. GC–MS chromatograms of plant extracts were investigated to test the applicability of the developed procedure to identify and quantify the strong odd/even predominance displayed by biogenic n-alkanes. In fact, it is known that vascular plants synthesize epicuticular waxes containing odd carbon-number n-alkanes.[12,14,37–39] The application is particularly relevant since identification and characterization of long-chain n-alkanes (C_{27}–C_{35} terms) from leaf wax, where they represent a minor portion of the overall wax composition, has proved to be the basis of a taxonomic system for classifying plant groups.[40]

Dichloromethane extracts of leaves and flowers of different plant families were submitted to GC–MS: the SIM chromatogram was monitored at $m/z = 57 + 71 + 85$ to represent the aliphatic hydrocarbon fraction. As an example, chromatograms of hydrocarbons extracted from 'Mimosa' flower are pictured in Fig. 4a.

The main components are mid- and long-chain n-alkanes C_{21}–C_{33}. C_{23}, C_{25}, C_{27} and C_{29} are the dominant long-chain n-alkanes in the GC profiles. A visual examination of the chromatogram shows a typical chromatographic profile of n-alkanes from vascular land plants characterized by a high odd-to-even predominance of long chain C_{25}–C_{33} with $CPI \approx 5$–10.

The $EACVF_{\mathrm{tot}}$ plot computed on the whole chromatogram (Fig. 4b) clearly shows a bimodal distribution of the $EACVF_{\mathrm{tot}}(bk)$ peak height with lower values at odd k values (combination term, eqn (8)) and higher values at even k (addition term, eqn (7)). Such a pattern is diagnostic of an odd/even prevalence that can be quantified by computing CPI_{tot} according to eqn (14): a CPI_{tot} value close to 5 was estimated (estimated values, 4th row in Table 2). To better characterize the plant chemical composition, the CPI_{bio} index was also computed by selecting the chromatographic region containing the long chain C_{24}–C_{33} terms (estimated values, 4th column in Table 2). The developed algorithm also yields an estimation of the n_{max} n-alkanes present in the sample (estimated values, 5th column in Table 2) directly from $EACVF_{\mathrm{tot}}$ peaks at $\Delta t = bk$ even k (eqn (16)).

The analysis of the *H. perforatum* sample permits a direct comparison of the present results with data in the literature since the *Hypericum* species has been intensely studied, in particular as regards the composition and abundance of n-alkanes.[33,34,41] The CPI_{bio} and n_{max} values estimated from the $EACVF_{\mathrm{tot}}$ plot ($CPI_{\mathrm{bio}} = 15$ and $n_{\mathrm{max}} = 7$, 6th row in Table 2) closely agree with

This journal is © The Royal Society of Chemistry 2009

Table 2 *CPI* and n_{max} parameters for experimental chromatograms of real samples. 1st–5th columns: estimation by using $EACVF_{tot}$ method; 6th–9th columns: calculation by traditional method

	$EACVF_{tot}$ estimation				Traditional calculation			
Sample	CPI_{tot}	CPI_{pet}	CPI_{bio}	n_{max}	CPI_{tot}	CPI_{pet}	CPI_{bio}	n_{max}
Fuel	1.60	1.67	—	19.4	0.97	0.96	—	20
Fuel (TIC)	1.52	1.56	—	19.0	1.03	1.00	—	20
ASTM D2887	1.35	1.56	—	22.6	1.02	1.08	—	22
Mimosa	5.30	—	5.80	12.4	5.80	—	4.67	11
Chrysanthemum	7.67	—	5.85	13.2	7.88	—	6.34	13
Hypericum	14.6	—	14.6	7.17	12.4	—	12.4	8

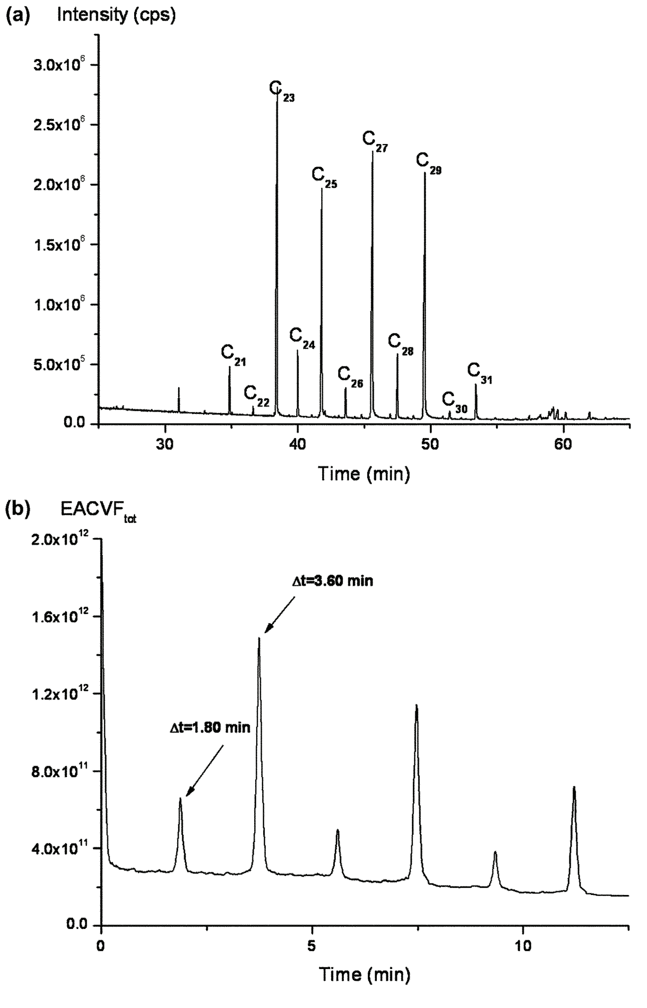


Fig. 4 Dichloromethane extracts of 'Mimosa' flowers. (a): SIM signal of the GC–MS chromatogram monitored at $m/z = 57 + 71 + 85$; (b): plot of $EACVF_{tot}$ computed on the whole chromatogram: the arrows identify the peaks diagnostic of the sequence at $\Delta t = b = 1.80$ min and $\Delta t = 2b = 3.60$ min.

literature ($CPI_{bio} = 15$ and $n_{max} = 7$[41]). It must be underlined that these CPI_{bio} values are considerably higher than in other plant species and are due in part to the high percentage of the C_{29} term.[15]

To check the accuracy of the obtained results, the traditional procedure based on calculation on each identified and integrated peak was applied to compute CPI_{tot}, CPI_{bio} and n_{max} values (traditional calculations, 6th–9th columns in Table 2). The close similarity between the computed and the estimated $EACVF_{tot}$ data proves the reliability of the developed method to identify and characterize the abundance distribution of biogenic n-alkanes, and this may also be useful in extracting information for a chemotaxonomic approach.

Conclusions

A method was developed and validated for processing and calculating data from the analysis of complex chemical mixtures that otherwise would be cumbersome and time-consuming. The procedure is specifically devoted to handling the extensive amounts of data generated by the hyphenated chromatographic techniques when applied to complex mixtures of contaminants, as those present in polluted environmental samples (*e.g.* sediment, soil, sludge, and biota).

The procedure is focused on a description of the chemical pattern of n-alkane homologous series, in particular, on the reliable computation of the CPI_{tot} index, as a descriptor of characteristic n-alkane distributions to be used as a signature of specific organic sources. It has proved suitable for the study of long chain n-alkane distributions dominated by odd carbon-numbered homologs, reliable indicators of terrigenous inputs in environmental and paleo-environmental studies. In addition, the method may be applicable for diagnostic fingerprinting ratios in relation to forensic oil spill identification or bioindicating of the general degree of environmental pollution.

The method may be extended to different homologous series in order to characterize specific organic markers for identifying sources and tracing inputs in the environment. The results directly obtained by computing $EACVF_{tot}$ on the whole chromatographic signal can constitute the basis for further data analysis using multivariate statistical methods, such as discriminant analysis (DA), cluster analysis (CA) or principal component analysis (PCA) to gain a better understanding the organic component contribution of inputs in the environment.

Acknowledgements

This work has been supported by the Italian University and Scientific Research Ministry (grant #2005032388 004), and by the University of Ferrara, Italy.

References

1 P. A. Cranwell, *Org. Geochem.*, 1981, **3**, 79–89.
2 E. D. Evans and E. E. Bray, *Geochim. Cosmochim. Acta*, 1961, **22**, 2–15.

3 G. Rieley, R. J. Collier, D. M. Jones and G. Eglinton, *Org. Geochem.*, 1991, **17**, 901–912.
4 J. I. Hedges and F. G. Prahl, in *Organic Geochemistry: Principles and Applications*, ed. M. H. Engel and S. A. Macko, Plenum Press, New York, 1993, vol. 1, pp. 237–253.
5 R. P. Eganhouse and I. R. Kaplan, *Environ. Sci. Technol.*, 1982, **16**, 180–186.
6 M. Brault, B. R. T. Simoneit, A. Saliot and J. C. Marty, *Org. Chem.*, 1988, **12**, 209–219.
7 K. Pendoley, *Mar. Pollut. Bull.*, 1992, **24**, 210–215.
8 W. Jeng, *Mar. Chem.*, 2006, **102**, 242–251.
9 M. C. Kennicutt II, C. Barker, J. M. Brooks, D. A. DeFreitas and G. H. Zhu, *Org. Geochem.*, 1987, **11**, 41–51.
10 W. C. Qu, B. Xue, C. W. Su and S. M. Wang, *Hydrobiologia*, 2007, **581**, 89–95.
11 M. Obermajer, K. G. Osadetz, M. G. Fowler and L. R. Snowdon, *Org. Geochem.*, 2000, **31**, 959–976.
12 C. M. Reddy, T. I. Eglinton, R. Palić, B. C. Benitez-Nleson, G. Stojanović, I. Palić, T. I. Djordjevć and G. Eglinton, *Org. Geochem.*, 2000, **31**, 331–336.
13 M. Maffei, S. Badino and S. Bossi, *J. Biol. Res.*, 2004, **1**, 3–19.
14 G. Eglinton and R. J. Hamilton, *Leaf Epicuticular Waxes Sci.*, 1967, **156**, 1322–1335.
15 M. A. Sonibare, A. A. Jayeola and A. Egunyomi, *Biochem. Syst. Ecol.*, 2005, **33**, 79–86.
16 J. Bendle, K. Kawamura, K. Yamazaki and T. Niwai, *Geochim. Cosmochim. Acta*, 2007, **71**, 5934–5955.
17 I. G. Kovouras and E. G. Stephanou, *Indoor Air*, 2002, **12**, 17–32.
18 A. Cincinelli, M. Del Bubba, T. Martellini, A. Gambaro and L. Lepri, *Chemosphere*, 2007, **68**, 472–478.
19 G. Wang, L. Huang, X. Zhao, H. Niu and Z. Dai, *Atmos. Res.*, 2006, **81**, 54–66.
20 Z. Wang and M. F. Fingas, *Mar. Pollut. Bull.*, 2003, **47**, 423–452.
21 L. Xu, L.-J. Tang, C.-B. Cai, H.-L. Wu, G.-L. Shen, R.-Q. Yu and J.-H. Jiang, *Anal. Chim. Acta*, 2008, **613**, 121–134.
22 M. Vosough and A. Salemi, *Talanta*, 2007, **73**, 30–36.
23 A. Felinger, L. Pasti and F. Dondi, *Anal. Chem.*, 1990, **62**, 1846–1853.
24 F. Dondi, A. Betti, L. Pasti, M. C. Pietrogrande and A. Felinger, *Anal. Chem.*, 1993, **65**, 2209–2215.
25 M. C. Pietrogrande, M. Mercuriali and L. Pasti, *Anal. Chim. Acta*, 2007, **594**, 128–138.
26 M. C. Pietrogrande, M. G. Zampolli and F. Dondi, *Anal. Chem.*, 2006, **78**, 2576–2592.
27 M. C. Pietrogrande, M. G. Zampolli, F. Dondi, C. Szopa, R. Sternberg, A. Buch and F. Raulin, *J. Chromatogr., A*, 2005, **1071**, 255–261.
28 A. Felinger, L. Pasti, P. Reschiglian and F. Dondi, *Anal. Chem.*, 1990, **62**, 1854–1860.
29 M. C. Pietrogrande, F. Dondi, A. Felinger and J. M. Davis, *Chemom. Intell. Lab. Syst.*, 1995, **28**, 239–258.
30 F. Dondi, M. C. Pietrogrande and A. Felinger, *Chromatographia*, 1997, **45**, 435–440.
31 M. C. Pietrogrande, I. Tellini, A. Felinger, F. Dondi, C. Szopa, R. Sternberg and C. Vidal-Madjar, *J. Sep. Sci.*, 2003, **26**, 569–577.
32 S. Pignatti, in *Flora d'Italia*, Edagricole, Bologna, 1st edn, 1982, vol. 1, pp. 1–3.
33 F. Conforti, G. A. Statti, R. Tundis, A. Bianchi, C. Agrimonti, G. Sacchetti, E. Andreotti, F. Menichini and F. Poli, *Nat. Prod. Res.*, 2005, **19**, 295–303.
34 J. Barnes, L. A. Anderson and J. D. Phillipson, *J. Pharm. Pharmacol.*, 2001, **53**, 583–600.
35 G. A. Caravaggio, J.-P. Charland, P. Macdonald and L. Graham, *Environ. Sci. Technol.*, 2007, **41**, 3697–3701.
36 M. Li, S. R. McDow, D. J. Tollerud and M. A. Mazurek, *Atmos. Environ.*, 2006, **40**, 2260–2273.
37 A. G. Douglas and G. Eglinton, in *The Distribution of Alkanes, Comparative Phytochemistry*, ed. T. Swain, Academic Press, London, 1966, pp. 57–78.
38 P. E. Kolattukudy, *Plant Waxes Lipids*, 1969, **5**, 259–275.
39 Z. Zhang, M. Zhao, X. Yang, S. Wang, X. Jiang, F. Oldfield and G. Eglinton, *Org. Geochem.*, 2004, **35**, 595–613.
40 M. A. Sonibare, M. O. Soladoye and Y. Ekine-Ogunlana, *Afr. J. Biotechnol.*, 2007, **13**, 1516–1520.
41 G. Stojanovic, R. Palic, C. H. Tarr, C. M. Reddy and O. Markinovic, *Biochem. Syst. Ecol.*, 2003, **31**, 223–226.

This journal is © The Royal Society of Chemistry 2009

PAPER IV

GC-MS analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosol: comparison between silylation and esterification derivatization procedures

Anal Bioanal Chem (2010) 396:877–885
DOI 10.1007/s00216-009-3212-z

ORIGINAL PAPER

GC–MS analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosol: comparison between silylation and esterification derivatization procedures

Maria Chiara Pietrogrande · Dimitri Bacco · Mattia Mercuriali

Received: 17 July 2009 / Revised: 29 September 2009 / Accepted: 4 October 2009 / Published online: 22 October 2009
© Springer-Verlag 2009

Abstract This paper describes methods for the determination of low-molecular-weight (LMW) dicarboxylic acids in atmospheric aerosols as important chemical tracers for source apportionment of aerosol organics and for studying atmospheric processes leading to secondary organic aerosol formation. The two derivatization procedures most widely used in GC analysis of dicarboxylic acids were compared: esterification using BF_3/alcohol reagent and silylation using *N*,*O*-bis(trimethylsilyl)-trifluoroacetamide (BSTFA). The advantages and drawbacks of the two methods are investigated and compared in terms of (1) precision and accuracy of the results and (2) sensitivity and detection limit of the procedure. The comparative investigation was performed on standard solutions containing target C_3–C_9 dicarboxylic acids and on experimental particulate matter (PM) samples. Attention was focused on low-volume sampling devices that collect small amounts of sample for organic speciation. The results show that, overall, both the techniques appear suitable for the analysis of LMW dicarboxylic acids in atmospheric aerosols since they provide low detection limits ($\leq$4 ng m^{-3}) and satisfactory reproducibility (RSD%$\leq$15%). Between them, BSTFA should be the reagent of choice under the most limiting conditions of PM filters collected by low-volume air samplers: It provides determination of all the target C_3–C_9 dicarboxylic acids with lower detection limits ($\leq$2 ng m^{-3}) and higher reproducibility (RSD%$\leq$10%)

Keywords Low-molecular-weight dicarboxylic acids · Derivatization procedures · GC–MS analysis · Atmospheric aerosol · Chemical marker

Introduction

Dicarboxylic acids are an important group of water-soluble organic compounds. They are ubiquitous in the troposphere and represent a significant fraction of atmospheric organic particulate matter (PM): Total diacids account for about 1–3% of the total particulate carbon in urban areas and even more than 10% in the remote marine environment [1–4]. They have received a great deal of attention because of their potential effect on the global climate: Diacids can modify the surface tension and hygroscopic properties of atmospheric particles, owing to their high water solubility and low vapor pressure [5]. Dicarboxylic acids have been extensively measured in total suspended particulate and PM_{10} samples collected in urban sites [2–4, 6–10] and continental background [11–17].

Detailed investigations have demonstrated that the concentration and relative abundance of these acids are useful organic tracers for source apportionment and atmospheric processes leading to secondary organic aerosol formation [2, 8, 18, 19]. In fact, such polar organics are emitted directly into the atmosphere as PM by a multiplicity of different sources and are also produced through secondary organic aerosol formation initiated by photochemical reactions incurred as ozone, hydroxyl, and nitrate radicals react with volatile hydrocarbons [1–4, 6]. In particular, low-molecular-weight dicarboxylic acids (C_3–C_9) may yield relevant information on the source strength of anthropogenic vs. biogenic precursors [2–4, 9, 20]. It has been suggested that the C_3/C_4 ratio is an indicator of

M. C. Pietrogrande (✉) · D. Bacco · M. Mercuriali
Department of Chemistry, University of Ferrara,
Via L. Borsari, 46,
44100 Ferrara, Italy
e-mail: mpc@unife.it

enhanced photochemical production of dicarboxylic acids in the atmosphere since succinic acid (C_4) is a precursor of oxalic (C_2) and malonic (C_3) acids. On the other hand, the C_6/C_9 ratio has been used as an indicator of the relative source strength of anthropogenic and biogenic diacid precursors: Adipic acid was proposed as a product of the oxidation of anthropogenic cyclohexene, while azelaic acid was thought to come from the oxidation of biogenic unsaturated fatty acids [3, 6, 9, 20].

To date, GC–MS is the one of the methods of choice for characterizing individual organic compounds within aerosol samples, primarily because of its high sensitivity and resolving power. The high polarity and low levels (approximately 1 ng m^{-3}) of dicarboxylic acids demand a derivatization step prior to GC analysis to reduce the polarity of the compounds. The most common reactions used to modify compounds containing acidic hydrogens are alkylation, acylation, and silylation [2–4, 6–16, 20–23]; among them two derivatization processes are mainly used to analyze dicarboxylic acids in PM samples because they offer easy sample preparation and display good analytical characteristics:

1. Esterification of the acid groups using methanol or 1-butanol as derivatizing agent in the presence of a relatively strong acid (BF_3 or BCl_3) [3, 6, 7, 9, 10, 13, 21] (first applied by Kawamura and co-workers [24]).
2. Silylation based on silylation reagents to form trimethylsilyl (TMS) derivatives [2, 3, 14–16, 21, 23, 25, 26].

The two methodologies differ in terms of the stability of the derivatives formed, the presence of interfering by-products, and speed. Moreover, a combination of the two procedures has been employed to yield a multistep derivatization by which –COOH groups are initially derivatized with BF_3/alcohol, and then the remaining hydroxy or keto groups are silylated with a silylation reagent [21].

BF_3 esterification

The BF_3/alcohol reagent converts either carboxyl groups into butyl esters or aldehyde groups into dibutyl acetals [17, 18]. Starting from the original Kawamura paper [24], different modifications have been reported and widely applied to make BF_3/alcohol derivatization the most widely used procedure for determining low-molecular-weight (LMW) oxygenates in atmospheric samples [3, 6, 7, 9, 10, 13, 21]. In particular the BF_3/butanol procedure has distinct advantages for quantifying LMW compounds because the resulting butyl derivatives are less volatile and more resistant to evaporative losses than the BF_3/methanol scheme [7, 10]. Because of the presence of residual acid, the products cannot be directly injected into the GC, but rather a purification step is required before injection [20].

The electron impact (EI) ionization of the butyl derivatives yields mass spectra including some common fragment ions $m/z=57$ ($[C_4H_9]^+$), $m/z=41$ ($[CH_2CH{=}CH_2]$) and $m/z=73$ ($[-OC_4H_9]^+$) arising from $-OC_4H_9$ moiety [21]. Common fragmentation pathways are also the cleavage of the C–O bond adjacent to the butyl group, which gives rise to the $[M-73]^+$ fragment, and the additional loss of an alkene fragment: They give rise to an $[M-129]^+$ ion fragment, which is the base peak ion for most C_3–C_9 dicarboxylic acids.

Silylation

The silylation reaction converts the hydroxyl groups into their corresponding trimethylsilyl derivatives via a substitution reaction, which yields one main product for each compound and with high conversion efficiency [23, 26]. The reagents commonly used for PM analysis are trimethylchlorosilane (TMCS), *N*-methyl-trimethylsilyltrifluoroacetamide, *N*,*O*-bis-(trimethylsilyl)trifluoroacetamide (BSTFA) and *N*-(*t*-butyldimethylsilyl)-*N*-methyltrifluoroacetamide (MTBSTFA) [2, 3, 14–16, 21, 23, 25, 26]. MTBSTFA should be preferred since its derivatives display a simplified fragmentation pattern yielding fragments with very high relative abundances, especially for $[M-57]^+$, that generates good detection limits. However, it was found that steric hindrance and molecular mass play a very important role in the choice of the best suited derivatization reagent: MTBSTFA derivatization of compounds with sterically hindered sites produces very small analytical responses or no signal at all, and BSTFA derivatization of compounds with high molecular mass produces no characteristic fragmentation pattern. Therefore, the use of BSTFA is the best choice for analysis of sugars or saccharides, which are another class of polar organics commonly determined in PM samples as molecular tracers in elucidating organic carbon sources and atmospheric transport pathways [23]. For these reasons, the present study describes the use of BSTFA for a comprehensive procedure that can be extended to analysis of a wider range of polar organics including sugars.

The BSTFA reaction is moisture sensitive and requires mild conditions to complete the derivatization in order to achieve GC–MS detection at very low concentrations [23, 25, 26]. In opposition to alkylation, silylation normally does not require a purification step, and the derivatives can be injected directly into the GC system [20, 23, 25, 26]. However, it presents some drawbacks, such as the fact that the silylation reagent is dangerous and some artifacts can be produced in the reaction [21, 25].

In the EI mass spectra of TMCS compounds rearrangement reactions of the trimethylsilyl group may occur, making EI mass spectra quite complex and difficult to interpret. Generally, the BSTFA derivatives display a common fragmentation pattern formed by ion fragments at *m/z*=73 and 75, $[Si(CH_3)_3]^+$ and $[HO{=}Si(CH_3)_2]^+$, respectively, derived by substitution of the active H atom with the $–Si(CH_3)_3$ group. In addition, compounds with two active H atoms, such as dicarboxylic acids, show abundant ions with *m/z*=147, postulated as $[(CH_3)_2Si{=}Si(CH_3)_2]^+$, and this is accompanied by *m/z*=149 ion resulting from the hydrogenation of *m/z*=147 that occurs in the ion trap [23, 26].

This paper focuses on the determination of LMW dicarboxylic acids (C_3–C_9) because they contain relevant chemical information to distinguish primary vs. secondary sources as well as anthropogenic vs. biogenic precursors. Preference was given to a faster one-step derivatization procedure to determine selected target compounds: The advantages and drawbacks of the methods using BF_3/alcohol and BSTFA are investigated and compared in terms of precision and accuracy of the results, sensitivity, and detection limit of the procedure.

Experimental

Reagents and standards

Reagents used for the different derivatizations (BF_3-1-butanol and BSTFA 1% trimethylchlorosilane) were obtained from Aldrich Chemical Co. (Milan, Italy). All standards and reagents used were of the highest purity commercially available. Dicarboxylic acid standards were purchased from Fluka/Aldrich/Sigma (Sigma Aldrich, Srl, Milan, Italy). All solvents were trace analysis grade (from 99.7%) from Sigma Aldrich (Milan, Italy).

Individual stock standard solutions were prepared in methanol for each C_3–C_9 dicarboxylic acid at concentrations varying from 500 to 1,000 µg L^{-1}. These solutions were diluted serially—using water obtained from a Milli-Q water purification system (Millipore, Vimodrone, Milan, Italy)—to prepare lower concentration solutions to compute calibration curves and assess acid recoveries (proper concentration to obtain an absolute injected quantity ranging from 1.7 to 28 ng of each acid). These quantity values were also translated into air volume concentrations (1.5–25 ng m^{-3}) by assuming the analysis of one filter collected over 24 h by a low-volume sampler (55 m^3 air volume). The two derivatizing agents, as well as the individual and composite standard solutions, were stored at 4 °C.

Extraction of environmental samples

The PM_{10} samples were collected on a precombusted quartz fiber filter (20×25 cm) with an automatic outdoor station consisting of a low-volume sampler (Skypost PM, TCRTECORA Instruments, Corsico, Milan, Italy) operating at a flow rate of 38.3 L min^{-1} for 24 h. The samples were collected in a rural area (San Pietro Capofiume, Bologna, Italy) in spring (April 2008). After sampling, the procedure outlined in European Standard EN 12341 (CEN, 1998) was applied for equilibration and weighing.

Filter samples were extracted for 30 min in an ultrasonication bath with pure Milli-Q water (3×10 mL), and then the extract aliquots were combined and filtered using a glass fiber filter (42.5 mm, GF Grade, Whatman, Maidstone, UK) to remove insoluble particles. The filtrate was then evaporated completely using a stream of high-purity nitrogen.

Procedural blanks were run in order to monitor significant background interferences in environmental samples.

Derivatization procedures

BF_3 esterification

The procedure used is based on a modification [7] of the Kawamura method [24]. Standard solutions of the acids were put into a pear-shaped flask and evaporated to dryness in a gentle nitrogen stream at room temperature. A 14% BF_3–butanol (10 µL) mixture and 0.2 mL of *n*-hexane were added to the samples: The flask was capped with a ground-glass stopper and clamp and sealed with Teflon tape. The sample was heated for 60 min at 70 °C and then cooled to room temperature. During the reaction, dicarboxylic acids were converted into their corresponding butyl esters. After, 0.2 mL of water saturated with sodium chloride was added to neutralize the BF_3 excess, and the solution was allowed to stand for 2 min. To extract the sample for analysis, 0.5 mL of *n*-hexane was added to the sample tube, and then the tube was capped tightly and shaken vigorously for 3 min. The organic layer was transferred into a 2.5-mL tube and reduced to dryness under high-purity nitrogen stream. Finally, 100 µL of *n*-hexane and 5 µL of *n*-hexadecane solution (at 0.5 ng µL^{-1}, as an injection internal standard (IS)) were added, and then 2 µL of the sample was injected into the GC system.

Silylation

Derivatization of the dicarboxylic acids using BSTFA was performed following the procedure reported in detail elsewhere [26]. The sample was transferred into a 2.5-mL tube, and the solution evaporated to dryness; then 10 µL of

BSTFA plus 1% TMCS and 85 µL of isooctane were added to form TMS derivatives; 5 µL of *n*-hexadecane solution was added as an injection IS. The tube was sealed with a Teflon-coated cap, and the reaction was performed at 75 °C for 90 min. Then 2 µL of the sample was injected into the GC–MS system.

GC–MS analysis

The GC–MS system was a Scientific Focus-GC (Thermo-Fisher Scientific, Milan, Italy) coupled with PolarisQ Ion Trap Mass Spectrometer (Thermo-Fisher Scientific, Milan, Italy). The column used was a DB-5 column (L=30 m, I.D.=0.25 mm, df=0.25 µm film thickness; J&W Scientific, Rancho Cordova, CA, USA). High-purity helium was the carrier gas with a velocity of 1.5 mL min^{-1}.

Temperature program conditions were optimized for analysis of butyl and silyl derivatives. For butyl derivative analysis the column temperature program consisted of an initial isothermal step at 70 °C for 5 min, a temperature increase to 160 °C at 5 °C min^{-1}, followed by another increase to 280 °C at 15 °C min^{-1}. For silyl derivatives the temperature program started with an initial temperature of 75 °C (hold for 5 min); it was raised to 135 °C at 2 °C min^{-1}, followed by an isothermal hold for 2 min, and after that, the temperature was increased to 160 °C at 2 °C min^{-1}, then further raised to 280 °C at 15 °C min^{-1}.

All samples were injected in splitless mode (splitless time, 30 s); the injector temperature was 250 °C.

The mass spectrometer operated in EI mode (positive ion, 70 eV). Ion source and transfer-line temperatures were 270 and 320 °C, respectively. The mass spectra were acquired with repetitive scanning from 50 to 600 m/z in 1 s. The full scan detection mode was chosen since it allows a comprehensive investigation of the wide range of polar organics yielding derivatives under the selected operative conditions. In addition to total ion chromatograms (TIC), the selected-ion monitoring (SIM) mode was used to quantify the target analytes: Either the base peak ion or one of the most abundant characteristic fragments was chosen as the SIM ions (Table 1, third and sixth columns).

Compound identification was performed by comparison of the chromatographic retention times and mass spectra with those of authentic standards and the mass spectral library of the GC–MS data system.

All samples were analyzed in triplicate. To obtain reliable and reproducible quantitative data, the internal standard procedure was used. Hexadecane was added as an injection IS since it is not subject to the derivatization procedures: The detector response was expressed as peak area value, A_{ca}, relative to internal standard peak area (A_{IS}), i.e., A_{ca}/A_{IS}.

Analytical parameters of the GC–MS method

The method sensitivity and linearity were evaluated by computing the calibration curves using six multicomponent standard solutions containing C_3–C_9 acids in a concentration range corresponding to 1.5–25 ng m^{-3} in the sampled air (absolute injected quantity 1.7–28 ng) for each compound. Samples were derivatized and analyzed by GC–MS: Each point on the curve, obtained as the average of three replicated measurements, is expressed as peak area ratio A_{ca}/A_{IS}.

The regression parameters were computed by the least-squares method: The intercept values were verified as statistically equal to zero by computing the 95% two-sided confidence interval for each intercept value (b_0) and applying the t test at 5% of significance.

Sensitivity was assessed by establishing the detection limit X_{LOD} for each studied acid. On the basis of the slope of the calibration line, X_{LOD} was computed as the analyte concentration yielding a signal value $Y_{LOD} = y_b + 6\sigma_b$, where y_b is the blank average signal of 10 blank responses, and σ_b its standard deviation. This X_{LOD} value corresponds to a 0.13% probability that the blank signal will be misinterpreted and that the compound may be lost [27]. The detection limit was computed as $X_{LOD} = 6\sigma_b/b_1$, where b_1 is the slope of the calibration line.

To check the precision and accuracy of the proposed method, recovery experiments were carried out by spiking known amounts of target C_3–C_9 dicarboxylic acids onto one half of a blank quartz fiber filter and processing the spiked filters as the real aerosol samples. Three different concentration levels were investigated for each analyte, i.e., 10, 20, and 30 ng m^{-3}, and the measurements were repeated in triplicate to compute mean recovery and RSD% values.

Results and discussion

The two most common derivatization procedures were compared for quantitative analysis of dicarboxylic acids in PM samples by focusing attention on two challenging conditions:

1. Quantification of lighter C_3 and C_4 dicarboxylic acids, since they contain relevant information for source apportionment and secondary organic aerosol formation [2–4, 6, 7, 9, 20].
2. Analysis of PM samples collected by low-volume devices (55 m^3 air volume sampled over 24 h) requiring the highest method sensitivity at trace level [4, 8, 14, 17, 21].

BF_3 esterification

After derivatization according to the described procedure ("Extraction of environmental samples"), the standard

Table 1 Parameters for GC–MS analysis of the target C_3–C_9 dicarboxylic acids using esterification (first to third columns) and silylation (fourth to sixth columns) procedures

Acids	tr (min)	*m*/*z* max	*m*/*z* SIM	tr (min)	*m*/*z* max	*m*/*z* SIM
	BF_3 esterification			Silylation		
Malonic acid	–	–	–	14.4	*147*; 149	75; 147; 149
Succinic acid	–	–	–	20.6	*147*; 149	75; 147; 149
Glutaric acid	23.7	87; *115*	115	26.0	*147*; 149	75; 147; 149
Adipic acid	25.2	*111*; 129	129	31.8	*75*, 141	75; 147; 149
Pimelic acid	26.2	*125*; 143	143	37.6	75, *155*	75; 147; 149
Suberic acid	27.1	139; *157*	157	43.7	*75*; 149	75; 147; 149
Azelaic acid	27.9	125; *171*	171	49.4	75; *149*	75; 147; 149

Retention times of the analyte derivatives are in the first and fourth columns, major derivative fragment ions in the second and fifth columns (most intense fragments are in italics), and *m*/*z* values selected for SIM detection in the third and sixth columns

solutions were analyzed by EI in the full scan mode in order to investigate the fragmentation pattern of each compound. The $[M-129]^+$ ion fragment is the base peak ion for most butyl derivatives, consistently with Kawamura [21] (Table 1, second column). The *m*/*z* values selected for SIM detection and quantification were either the base peak ion or one of the more abundant characteristic fragment ions (Table 1, third column). Reliable measurements can be obtained only for the dicarboxylic acids heavier than C_5, since the lighter C_3–C_4 acids yield butyl esters that are too volatile to avoid evaporative losses and too unstable to be accurately quantified [10, 20]. The calibration curves were computed for C_5–C_6 acids (Table 2, first to fifth rows): The obtained parameters show that the derivatization procedure displays good linearity and sensitivity, nearly independent of the acid molecular weight. The achieved detection limits are low enough to make the method compatible with environmental analysis: 2.6–4.9 ng m^{-3} in the sampled air (Table 2, fourth column) corresponding to 2.9–5.4 ng as absolute injected quantity (Table 2). The reported X_{LOD} values resulted higher than some of the data in the literature [1, 7]. This result is consistent with the unfavorable conditions of low sampled volume (55 m^3), which are mainly critical for lighter dicarboxylic acids.

Recovery experiments were carried out by spiking one half of a blank quartz fiber filter with known amount, i.e., 10, 20, and 30 ng m^{-3} of target C_5–C_9 dicarboxylic acids. The results for the 20 ng m^{-3} concentration are listed in Table 3, where the mean recovery and RSD% values are reported for triplicate measurements (Table 3). Recovery of the target compounds varied from 66% for glutaric acid to 120% for azelaic acid (Table 3, first column). The standard derivation values were lower than 15% for the studied acids, with the exception of the lighter glutaric acid (21%; Table 3, second column). As expected, the precision of the method decreases with the analyte concentration since for

Table 2 Calibration curve parameters for the C_3–C_9 acids and X_{LOD} values

Acids	b_1	b_0 (ngm^{-3})	R^2	X_{LOD} (ngm^{-3})	X_{LOD}(ng)
BF_3 esterification					
Glutaric acid	0.0112±0.0008	−0.01±0.01	0.992	4.9	5.4
Adipic acid	0.0117±0.0004	0.001±0.008	0.994	3.5	3.9
Pimelic acid	0.0110±0.0004	0.006±0.007	0.997	2.9	3.2
Suberic acid	0.0110±0.0005	0.01±0.01	0.995	2.6	2.9
Azelaic acid	0.0120±0.0006	0.02±0.01	0.997	2.9	3.2
Silylation					
Malonic acid	0.100±0.003	−0.11±0.05	0.997	2.6	2.9
Succinic acid	0.135±0.005	−0.09±0.06	0.997	1.9	2.1
Glutaric acid	0.081±0.003	−0.06±0.05	0.995	2.6	2.9
Adipic acid	0.036±0.001	−0.03±0.01	0.997	2.1	2.3
Pimelic acid	0.032±0.001	−0.03±0.02	0.995	2.7	3.0
Suberic acid	0.0296±0.0008	−0.03±0.01	0.997	2.4	2.6
Azelaic acid	0.023±0.001	−0.02±0.01	0.997	2.1	2.3

First to fifth rows, BF_3esterification reaction (SIM signals at different *m*/*z* values specific for each acid, see Table 1); sixth to 12th rows, silylation (SIM signal at $m/z=147$ for all the studied acids). X_{LOD} values are expressed as air volume concentration (ng m^{-3}, fourth column) and absolute injected quantity (ng, fifth column)

Table 3 Accuracy (recovery %) and reproducibility (RSD%) of esterification and silylation procedures estimated on triplicate measurements at 20 ng m^{-3} concentration level

	R%	RSD%	R%	RSD%
	BF_3 esterification		Silylation	
Malonic acid	–	–	78	6
Succinic acid	–	–	85	3
Glutaric acid	66	21	95	8
Adipic acid	88	12	99	5
Pimelic acid	90	9	107	9
Suberic acid	95	9	110	6
Azelaic acid	120	15	115	10

all the acids the relative standard derivation values range from 24% to 16% for the 10 ng m^{-3} level and from 6% to 14% for the 30 ng m^{-3} concentration.

Silylation reaction

Different experimental conditions have been widely applied to derivatize LMW oxygenate compounds for subsequent GC determination in PM samples [2, 4, 14–16, 21, 23]. Starting from the data in the literature, an optimization study was performed on the derivatization conditions that most affect analytical responses—i.e., reaction temperature and duration time—in order to develop a rapid, reproducible quantitative method for trace levels. The study was performed on standard aqueous solutions of C_5–C_9 dicarboxylic acids (each acid at 20 μg mL^{-1}).

The standard solutions were derivatized according to the procedure described in "Extraction of environmental samples" and submitted to GC analysis MS using SIM detection mode: The most abundant ions with $m/z=147$ and $m/z=149$ are selected for SIM detection to differentiate compounds bearing −COOH from other classes of organics (Table 1, fifth and sixth columns). The analytical responses were expressed as relative peak area A_{ca}/A_{IS} (hexadecane as internal standard).

The influence of reaction temperature and duration was tested carrying out the reaction at 50, 75, and 100 °C for 30, 60, and 90 min. Among the experimental conditions exploited, the reaction conditions of 75 °C for duration of 90 min yielded the best results and were chosen in the following study.

The analytical performance of the procedure was assessed by the calibration curves computed on standard solutions of C_3–C_9 acids. The obtained results show that the procedure allows quantification of all the studied acids with good sensitivity, since it achieves low X_{LOD} values independent of the acid's molecular weight, ranging in the concentration values from 1.9 to 2.7 ng m^{-3} in the sampled air (Table 2, fourth column) corresponding to 2.1–3.0 ng interval as absolute injected quantity (Table 2, fifth column).

The results obtained from studying procedure precision and accuracy are reported in Table 3 (recovery and RSD% values for the 20 ng m^{-3} concentration). Good recoveries were found for all the target compounds ranging from 78% for malonic acid to 115% for azelaic acid (Table 3, third column). The procedure also displays good reproducibility, as evaluated by RSD% values on three replicates, that range from 3% to 10% (Table 3, fourth column). Moreover, these properties were nearly constant for the 10 and 30 ng m^{-3} concentration levels.

Comparison between silyl and butyl ester derivatization

The obtained results confirm that both the methods are reproducible, trace-level procedures suitable for environmental monitoring of dicarboxylic acids. However, some differences can be singled out when the two procedures are compared for the challenging application of quantitative determination of lighter C_3–C_9 dicarboxylic acids at trace levels. These differences will be discussed below in terms of their relevance in the environmental measurements for PM monitoring.

- The silyl derivatives of dicarboxylic acids have higher molecular weights and are less volatile than the respective butyl derivatives. Thus, less time is required to elute the butyl derivatives, thus significantly reducing the duration of a chromatographic run: Only 30 min is the retention time required for the butyl ester of the most retained C_9 acid, while nearly 50 min is needed for the corresponding silyl derivative (Table 1, first and fourth columns). Nevertheless, the higher volatility yields some evaporative loss during the derivatization procedure thus decreasing analyte recovery.
- One merit of the trimethylsilylated compounds is that their EI mass spectra yield structurally characteristic ions (ion fragment at $m/z=74$, 147, 149), which make identification highly reliable. These ion fragments can be used in SIM detection mode for a simpler and more selective chromatographic signal. This advantage cannot be achieved by butyl derivatives, since their prominent fragmentation pathway commonly gives rise to the $[M-73]^+$ fragment as the most abundant fragment ion. Therefore, different m/z values must be specifically selected for SIM detection and quantification of each target acid (Table 1, third column).
- In general, the method sensitivity obtained with BSTFA derivatives was higher than the one with the butyl esters. This can be explained by the silylation reaction yield (i.e., degree of conversion) or by the

stability of the derivatives during handling. As a consequence, the silylation procedure displays higher sensitivity with lower X_{LOD} values for all the investigated C_3–C_9 dicarboxylic acids, compared to butyl esterification. On the other hand, the sensitivity of the BF_3/BuOH method strongly depends on the acid molecular weight: It is unreliable for the lower C_3–C_4 terms, and it significantly increases with the acid molecular weight to achieve detection limits comparable to those of silylation for the heavier C_7–C_9 acids. The low sensitivity for C_3–C_4 acids is also due to the concomitant higher volatility of their derivatives, which yields evaporative loss during the derivatization procedure.

Application to real PM samples

In order to confirm the findings obtained upon standard solutions, the two methods were applied to environmental PM matrices. To obtain comparable results on the same aerosol sample, each sample (PM samples 1 and 2 in Table 4) was obtained by two consecutive sets of samples (two quartz fiber filters) combined for extraction and then halved to separately perform derivatization prior to GC–MS analysis. The TIC chromatogram of the BF_3/BuOH derivatized sample (sample 1) is reported in Fig. 1a (derivative retention times are reported in Table 1, first column). The silyl derivatives obtained with BSTFA reagents on the same aerosol sample (sample 1) were analyzed under SIM detection mode at m/z=74, 147, 149 (SIM chromatogram in Fig. 1b; derivative retention times are reported in Table 1, fourth column).

The concentrations of the target dicarboxylic acids were measured with both the procedures using the calibration curves reported in Table 2: The obtained results are reported in Table 4 for both the samples.

As verified on standards, the lighter C_3 and C_4 acids escaped detection by the BF_3/BuOH method. It must be noted that some target acids are present in the investigated samples at a concentration level close to their detection limit, in particular in the case of BF_3 esterification, where glutaric acid concentration is lower than X_{LOD} (Table 4). However, a good agreement within 4% was shown by the results obtained from the two procedures for all the quantified acids: This proves that both derivatization procedures produce reproducible quantification.

The individual species show similar abundance independent of the carbon chain length, with malonic and azelaic acids predominant. These results are consistent with literature on dicarboxylic acids in $PM_{2.5}$ for a rural sampling site [8, 13–17, 23]. The predominance of the C_9 diacid is expected since it is an oxidation product of biogenic unsaturated fatty acids. This result is confirmed by a low value of 0.5 computed for the C_6/C_9 ratio for both samples and using both procedures to indicate a high biogenic input for aerosol diacids (Table 4) [1, 2, 4, 6, 9, 20].

Moreover, BSTFA derivatization also makes it possible to compute the C_3/C_4 ratio as another marker of diacid origin: Both samples yield a value of 1.3 as it is commonly observed in atmospheric aerosols with low anthropogenic sources (combustion of fossil fuels produces $C_3/C_4 \approx 0.3$) and reduced photo-induced secondary formation of dicarboxylic acids (that would yield higher $C_3/C_4 \geq 3$ values) [3, 4, 6, 9, 20].

Conclusions

Comparison of the two popular derivatization reactions shows that, on the whole, both techniques are suitable for the analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosols since they provide low detection limits ($\leq$4 ng m^{-3}) and satisfactory reproducibility (RSD %$\leq$15%).

The BSTFA procedure is preferable when the analysis is performed under the most challenging conditions concerning determination of lighter C_3–C_4 terms in PM

Table 4 Concentrations of the target dicarboxylic acids measured on two experimental aerosol samples (samples 1 and 2) after derivatization with BF_3/BuOH (first and second columns) and BSTFA (third and fourth columns) reagents

Acids	PM sample 1 (ngm^{-3}) Esterification	PM sample 2 (ngm^{-3})	PM sample 1 (ngm^{-3}) Silylation	PM sample 2 (ngm^{-3})
Malonic acid	–	–	5.2±1.4	5.4±1.5
Succinic acid	–	–	3.8±1.2	4.0±1.6
Glutaric acid	2.6±3.3	2.8±3.2	2.5±1.6	2.7±1.4
Adipic acid	3.1±2.0	3.3±3.8	3.0±1.2	3.2±1.6
Pimelic acid	2.7±1.8	3.0±2.6	2.6±1.8	2.9±1.4
Suberic acid	2.6±2.7	3.0±3.9	2.5±1.4	2.9±1.3
Azelaic acid	5.8±2.8	6.1±3.4	5.6 ±1.8	6.0±2.0

filters collected by low-volume air samplers: It provides lower detection limits (≤2 ng m^{-3}) and higher reproducibility (RSD%≤10%). Moreover, the use of BSTFA reagent can be extended to other the polar compounds, although containing sterically hindered sites, such as sugars which are relevant molecular markers of biogenic sources. It is evident that the demand for high sensitivity is less constrictive if high-volume sampling devices are used and enough material is collected for detailed organic speciation.

It must be underlined that both procedures require water evaporation: In particular, in the BF_3/BuOH derivatization process water facilitates the reverse derivatization reaction. The water evaporation step is time consuming and causes significant evaporative losses of the smaller, more volatile target compounds (e.g., malonic acid).

Further work is underway, changing extraction procedure and derivatization operating conditions in order to simultaneously determine additional compounds that are either more or less water soluble (e.g., sugars and larger mono- and dicarboxylic acids) or more volatile (e.g., nonanal). Solutions may also turn to simplified procedures that drop the water extraction and evaporation steps, i.e., mixing the derivatization reagents directly with filter substrates or using direct thermal desorption device coupled

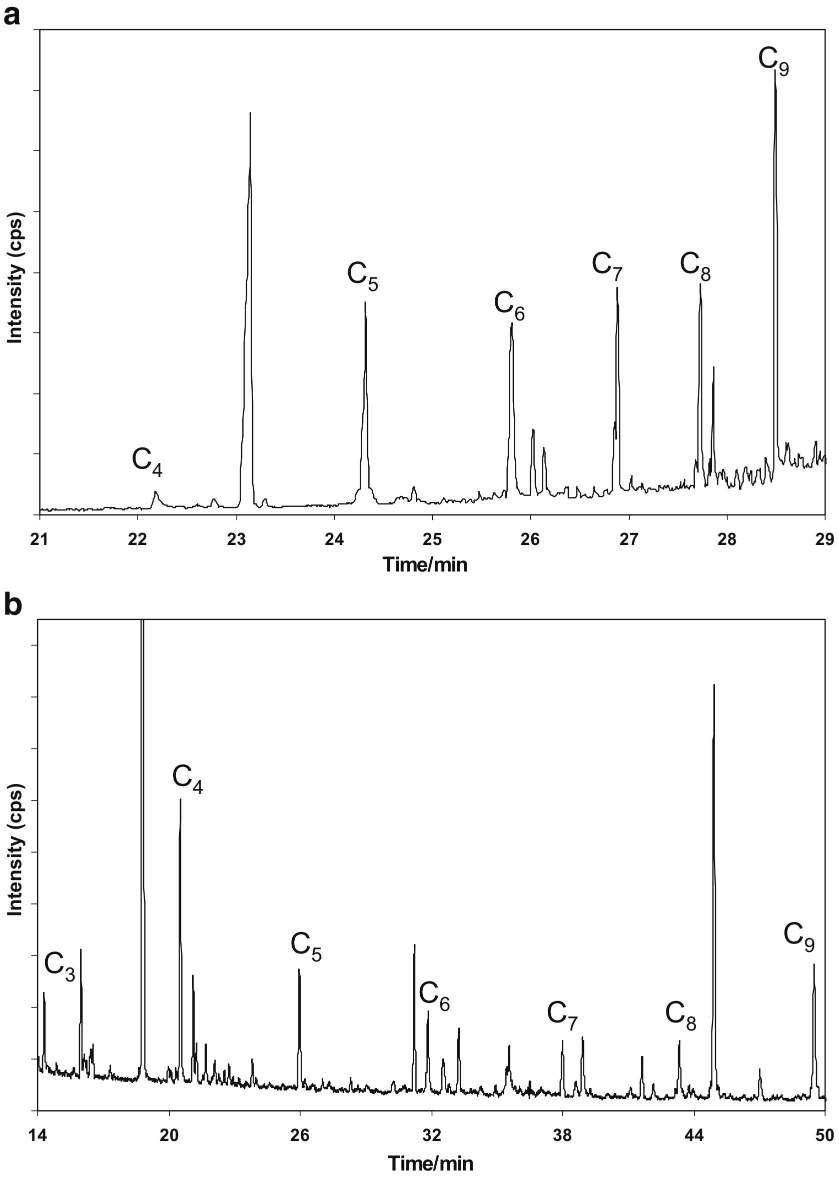


Fig. 1 GC–MS chromatograms of a derivatized environmental PM matrix (PM sample 1, in Table 4). **a** TIC chromatogram of the BF_3/BuOH derivatives; **b** SIM chromatogram at *m*/*z*=74, 147, 149 of the silyl derivatives

to GC–MS to reduce time-consuming sample preparation steps and analyte losses by desorption of the organics within the GC injector.

Acknowledgments The authors thank F. Dondi (Department of Chemistry, University of Ferrara) for the fruitful discussion and suggestions. This work was in part financially supported by the Moniter project of the Environment Agency of Emilia Romagna region (ARPAER). The authors thank the personnel of ARPAER for their support in providing the PM filters.

References

1. Collet JL Jr, Herckes P, Youngster S, Lee T (2008) Atm Res 87:232–241
2. Bi X, Simoneit BRT, Sheng G, Ma S, Fu J (2008) Atm Res 88:256–265
3. Kawamura K, Ikushima K (1993) Atm Env 27:2227–2235
4. Oliveira C, Pio C, Alves C, Evtyugina M, Santos P, Goncalves V, Nunes T, Silvestre AJD, Palmgren F, Wahlin P, Harrad S (2007) Atm Env 41:5555–5570
5. Peng C, Chan MN, Chan CK (2001) Environ Sci Technol 35:4495–4501
6. Wang H, Kawamura K, Ho KF, Lee SC (2006) Environ Sci Technol 40:6255–6260
7. Li YC, Yu JZ (2005) Environ Sci Technol 39:7616–7624
8. Li M, McDow SR, Tollerud DJ, Mazurek MA (2006) Atm Env 40:2260–2273
9. Ho KF, Lee SC, Cao JJ, Kawamura K, Watanabe T, Cheng Y, Chow JC (2006) Atm Env 40:3030–3040
10. Huang XF, He LY, Hu M, Zhang YH (2006) Atm Env 40:2449–2458
11. Kourtchev I, Copolovici L, Claeys M, Maenhaut W (2009) Environ Sci Technol 13:4665–4671. doi:10.1021/es803055w
12. Mochida M, Kawabata A, Kawamura K, Hatsushika H, Yamazaki K (2003) J Geophys Res 108:4193–4196
13. Kawamura K, Imai Y, Barrie LA (2005) Atm Env 39: 599–614
14. Kourtchev I, Warnke J, Maenhaut W, Hoffmann T, Claeys M (2008) Chemosphere 73:1308–1314
15. Yue Z, Fraser MP (2004) Atm Env 38:3253–3261
16. Lewandowski M, Jaoui M, Kleindienst TE, Offenberg JH, Edney EO (2007) Atm Env 41:4073–4083
17. Yu LE, Shulman ML, Kopperud R (2005) Environ Sci Technol 39:707–715
18. Schauer JJ, Rogge WF, Hildemann LM, Mazurek MA, Cass GR (2007) Atm Env 41:S241–S259
19. Shrivastava MK, Subramanian R, Rogge WF, Robinson AL (2007) Atm Env 41:9353–9369
20. Ray J, McDow SR (2005) Atm Env 39:7906–7919
21. Jaoui M, Kleindienst TE, Lewandowski M, Edney EO (2004) Anal Chem 76:4765–4778
22. Plewka A, Hofmann D, Müller K, Herrmann H (2003) Chromatographia 57:S253–S259
23. Schummer C, Delhomme O, Appenzeller BMR, Wennig R, Millet M (2009) Talanta 77:1473–1482
24. Kawamura K, Kaplan IR (1984) Anal Chem 56:1616–1620
25. Halket JM (1993) Handbook of derivatives for chromatography, 2nd edn. Wiley, New York, p 297
26. Pietrogrande MC, Zampolli MG, Dondi F, Szopa C, Sternberg R, Buch A, Raulin F (2005) J Chromatogr A 1071:255–261 Anal Chem, 2006, 78, 2579–2592
27. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi L, Smeyers-Verbeke J (1988) Handbook of chemometrics and qualimetrics: part A. Elsevier Science, Amsterdam, p 171

PAPER V

Distribution of n-alkanes in the northern Italy aerosols: data handling of GC-MS signals for homologous series characterization

Distribution of n-alkanes in the northern Italy aerosols: data handling of GC-MS signals for homologous series characterization

M. C. Pietrogrande*[1], M. Mercuriali[1], M.G. Perrone[2], L. Ferrero[2], G. Sangiorgi[2], E. Bolzacchini[2]

[1]Department of Chemistry, University of Ferrara, Via L. Borsari, 46, 44100 Ferrara, Italy;

[2]POLARIS Research Center, Department of Environmental Science, University of Milano-Bicocca, P.zza della Scienza 1, 20126 Milan, Italy

Correspondence: Prof. M. C. Pietrogrande, Department of Chemistry, University of Ferrara, Via L. Borsari, 46, I-44100 Ferrara, Italy. E-mail: mpc@unife.it

Keywords: n-alkanes, Carbon Preference Index, organic input source, chemical tracer

Abbreviations $ACVF_{tot}$, autocovariance function computed on the total chromatogram; $EACVF_{tot}$, experimental autocovariance function; SC, single component; CPI, Carbon Preference Index.

Abstaract

The paper describes the characterization of n-alkane homologous series present in PM samples performed by Gas Chromatography-Mass Spectrometry analysis. The PM samples were collected in three locations in northern Italy: Milan, a large urban area, Oasi Bine, a rural site far from big city centers, and Alpe San Colombano a remote, high altitude site in the Alps. They represent different particle sizes (PM_1, $PM_{2.5}$, PM_{10}) and seasons (summer, fall and winter). The analyzed samples were characterized in terms of PM total mass, total concentration of C_{20}-C_{32} n-alkanes and carbon preference index, CPI, to quantify the relative abundance of odd versus even n-alkanes.

As alternative to the conventional method based on peak integration, a chemometric approach based on computation of the Autocovariance Function ($EACVF$) was applied to extract homologous series property information. Two main parameters, CPI_{EACVF} and $series\%$, estimating the n-alkanes relative abundance, are derived from $EACVF$ and proved useful chemical markers for tracking the biogenic and anthropogenic origins of organic input sources.

The investigated samples display a large variation in the n-alkanes relative abundance: the lowest values ($series\% =$ 1-14%) were found in summer and the highest ($series\% =$24-48%) in winter, most likely the result of higher emissions from urban ‘‘winter’’ sources. In addition, a considerable seasonal variation in CPI_{EACVF} values can be identified for all the sampling sites. The CPI_{EACVF} values are close to 1 ($CPI_{EACVF} =$0.8-1.2) in the cold seasons, revealing a strong contribution from anthropogenic emissions, while the values show a greater spread ($CPI_{EACVF} =$0.9-3) in the warm season, i.e. reflecting the influence of contribution from biogenic sources.

The results obtained show the suitability of the $EACVF$ method in characterizing the n-alkane contribution and its applicability as a high-throughput method to analyze the huge amount of data derived from environmental monitoring. It increases result reliability by deconvolving complex signals into its components and reduces subjectivity of human intervention to increase data quality.

Introduction

Identification of the various sources of particulate matter (PM) and assessment of their chemical composition are important steps in air quality management. It has been recognized that atmospheric aerosol consists of a complex mixture of hundreds of compounds belonging to many different chemical classes [1-4]. Despite this complexity, in environmental monitoring and assessment studies, to adequately represent a chemical signature of the possible organic source inputs to atmospheric PM, attention is focused on chemical markers [5-8].

Homologous series of n-alkanes is a subgroup of the carbonaceous material especially suited to tracing the origin and fate of different samples: this is because they can originate from both man-made and natural sources and are highly resistant to biochemical degradation [9]. In particular, two parameters are of particular relevance in describing n-alkane properties as chemical signature of input sources: i) the number of terms and ii) the abundance distribution of the odd/even terms in the series. The latter property, introduced by Bray and Evans [10], is commonly expressed by the carbon preference index, CPI, which describes the relative abundance of odd- versus even-numbered carbon chain n-alkanes. It is a key diagnostic parameter in tracking the origin of organic inputs to determine the biogenic and anthropogenic nature of n-alkane sources. In particular, anthropogenic emissions from utilization of fossil fuel generate a random distribution of odd vs. even terms yielding CPI values close to 1. On the other hand, hydrocarbons originated from terrestrial plant material show a predominance of odd-numbered terms showing $CPI \approx 5-10$ [7,11-17].

Gas chromatography coupled with mass spectrometry (GC-MS) is the well-established technique of choice for identifying and quantifying the hydrocarbon fraction in complex mixtures of organics such as those present in aerosol samples [1,2,8]. The GC-MS signal obtained is usually a complex chromatogram, containing many resolved and unresolved peaks; for this reason, it is difficult to extract all the analytical information hidden in the chromatogram and hence the resulting estimate may be unreliable [18-20]. Moreover, the conventional chromatogram data processing method requires a great deal of labor and time since it needs to identify the n-alkane peaks by comparison to reference standards and MS spectra, integration of the identified peaks, computation of CPI as a ratio of the sum of concentrations of the odd vs. the even numbered carbon alkanes [1,2,8,16].

Therefore, signal processing procedures are very helpful in transforming the GC data into usable chemical information: in particular, a computer-assisted method has to be the preferred high-throughput approach since it reduces the labor and time required to handle the extensive amounts of data produced by environmental monitoring [18-21]. Among the many signal processing procedures developed to deal with this problem, a chemometric approach based on the AutoCovariance Function ($ACVF$) has been widely applied to experimental chromatograms [22-27]. Recently, an extension of the approach has been developed to extract information on the n-alkane series — n_{max} and CPI — directly from the $EACVF$ computed on the acquired chromatographic signal [17].

In this paper, this signal processing procedure is applied to handle GC-MS signals of PM samples collected at three sites in Italy, over different seasons (2008): thanks to the method, information on the relative contribution of the homologous series, $series\%$ and the CPI values, can be directly estimated from the $EACVF_{tot}$ and the PC computation takes just a few minutes.

GC-MS signal processing procedure based on Autocovariance Function. A chemometric approach based on the AutoCovariance Function ($ACVF$) has been developed to interpret the complex signals and the extensive amount of data obtained from GC-MS analysis. In particular, the method has proved powerful in extracting accurate information on the properties of the homologous series present in the analyzed mixture [17, 23-25].

The GC analysis of a sample containing terms of a homologous series — i.e. a PM sample containing n-alkane series — generates a signal formed by an ordered sequence of peaks whose retention time is described by:

$$t_R(n) = c + bn \qquad n = 0,1,2,3 \ldots n_{max} \tag{1}$$

where $t_R(n)$ is the retention time of the n^{th} term of the series; c the contribution a specific functional group makes to the overall retention; n_{max} the number of components belonging to the series. The value $\Delta t = b$ is the retention increment between the terms of the homologous series, e.g., the CH_2 retention time increment, in the strict case of linearized temperature programming of GC analysis [26]. A representative GC-MS signal of a PM sample is reported in fig. 1a (sample MI-17 of Table 1) where peaks of n-alkanes can be identified (arrows indicate the interdistance $\Delta t = b$).

The chemometric approach studies the Autocovariance Function (Experimental $ACVF$, $EACVF$) directly computed from the experimental chromatogram acquired in digitized form using the following expression [22]:

$$EACVF(\Delta t) = \frac{1}{M}\sum_{j=1}^{N-k} (Y_j - \hat{Y})(Y_{j+k} - \hat{Y}) \qquad k=0,1,2,\ldots M-1 \qquad (2)$$

where Y_j is the digitized chromatogram signal, $\hat{Y}$ its mean value, M the truncation point in the $EACVF$ computation. The correlation time $\Delta t = k\tau$ is the interdistance between the subsequent digitized positions, and assumes discrete values with k ranging from 0 to (M-1).

The $EACVF$ values represent the correlations between subsequent peaks in the chromatogram. They can be plotted as a function of the time interdistance Δt to obtain the $EACVF$ plot: a representative $EACVF$ plot computed on the GC-MS signal of fig. 1a is reported in fig. 1b.

A theoretical model has been developed and an algorithm has been implemented to estimate properties of the homologous series directly from the $EACVF$ values [17]. In particular, the following parameters can be obtained (detailed equations are reported in the Supporting Information):

1. *Presence of homologous series.* A sequence of GC peaks (following eq.1 due to terms of a homologous series) generates an $EACVF$ plot containing well defined deterministic peaks located at interdistance $\Delta t = b$ and multiple values, $\Delta t = bk$ (deterministic peaks indicated in fig. 1b). It can be shown that the $EACVF$ plot displays a simplified picture of the original signal, still retaining relevant diagnostic information able to reveal the presence of the series terms in the mixture. In fact, the power of the mathematic model lies in its ability to extract information on the ordered components, singling them out from the complexity of the signal.

2. *Number of terms of the homologous series.* The height of the $EACVF(bk)$ peaks computed at $\Delta t = bk$ for even k can be used to estimate the number, n_{max} (signed peaks in Fig 1b).

3. *Contribution of homologous series.* The ratio $EACVF(2b)/EACVF(0)$ can be considered an

estimation of the relative contribution of the homologous series terms to the total alkanes present in the sample. In fact, the $EACVF$ value computed at $\Delta t = 2b$ is related to the peak area of the homologous series components (arrow in Fig. 1b) while $EACVF(0)$ is related to the total area of all the chromatographic peaks, i.e., total alkane components, if the SIM signal at $m/z = 57+71+85$ is acquired.

4. *Odd/even prevalence***.** To quantitatively describe the abundance distribution of odd and even terms of the series, the R value is defined as the ratio:

$$R = \frac{a_{o,h}}{a_{e,h}} \tag{3}$$

where $a_{o,h}$ and $a_{e,h}$ are, respectively, the mean peak heights computed on the odd and even terms of the series. This R value can be properly used to estimate the CPI parameter, usually calculated on the same number of odd and even terms of the series. It is directly computed from the values of the $EACVF$ peaks at $\Delta t = b$ and $\Delta t = 2b$, since they retain information on the abundance of odd and even terms (see Supplementary Information).

5. Deconvolution of UCM band component. Experimental experience teaches that some PM samples — e.g., ambient aerosols with a high contribution from transportation or oil combustion — generate GC signals containing a large hump, known as the unresolved complex mixture (UCM), which interferes with the chromatographic separations of n-alkanes $\geq C_{24}$. This arises from the presence of other similar saturated nonpolar compounds, i.e., branched alkanes and alkylated cycloalkanes, which are difficult to be resolved with GC and which display similar fragmentation patterns under mass spectrometry [2,7,13,16] (a representative chromatogram is reported in Fig. 2a, sample MI-20 in Table 1).

The $EACVF$ plot computed on these signals shows the large UCM hump superimposed on the deterministic peaks at $\Delta t = kb$ diagnostic of the n-alkane sequence (Figure 2b, upper solid line). Nevertheless, the $EACVF$ method makes it possible to deconvolve the signal into the two components: the UCM ($EACVF_{UCM}$, dashed line in Fig 2b) and separated signal ($EACVF_{res}$, bottom bold line in Fig 2b) [26]. This is due to the intrinsic statistical property of the $EACVF$: i.e. it is additive when computed on independent variables such as the signals of UCM and resolved peaks [22]. The $EACVF_{res}$ plot shows well-defined deterministic peaks at $\Delta t = kb$ from which a reliable computation of CPI_{EACVF} parameter can be achieved.

As a simplified procedure, the height of the $EACVF$ peak at $\Delta t = 2b$, calculated from the peak baseline (bold arrows in Fig. 2b), can be proposed as an approximation of the $EACVF_{res}(2k)$ value and is used to estimate the relative abundance of alkane series ($series\%$). The accuracy of this approximation was verified by comparing the $EACVF$ results with the values computed using the traditional procedure based on computation performed on the integrated area of the identified n-alkane peaks.

Methods

Sample collection. The PM samples were collected at three locations in northern Itally. 38 samples were collected in Milan (MI; 45°31'19"N, 9°12'46"E) a large urban area: the sampling site was located at "Torre Sarca", where the University of Milano-Bicocca is located, an area with high motor vechicle traffic. Oasi Bine (OB; 45°08'40''N, 10°26'08''E) is located far from any big city centers, the nearest cities of Cremona and Mantova are about 15-20 km away: it represents the rural environment (22 sampled PMs). Other filters (18 samples) were collected at Alpe San Colombano (ASC, 2280 m. a.s.l; 46°27'18''N 10°18'50''E), a remote, high altitude site in the Alps.

At all the sampling sites, PM was collected by using low volume gravimetric samplers (flow 38.33 l/min: HYDRA sampler, FAI Instruments, Rome Italy). The samplers had two channels, each equipped, respectively, with a 2.5 μm and a 1 μm cutpoint inlet so that the $PM_{2.5}$ and PM_1 were collected daily (24 h), simultaneously on two Teflon filters (47 mm Ø, 2 µm, Pall Gelman, USA). PM_1 and $PM_{2.5}$ were sampled in MI, OB and ASC during the summer (S: June-August 08), fall (F: November 08: only MI and OB) and winter 2008-2009 (W: December 08-Janaury 09). In MI a 10 μm cutpoint inlet was also used to collect some PM_{10} samples during the summer and winter campaigns. A list of the studied samples is reported in Table 1: samples were categorized according the sampling site (1st column), particle size (2nd column) and seasonality (3rd column). Table 1 contains only the first 22 analyzed samples (from MI-1 to MI-22); for the complete data set of the analyzed samples, see the Supporting Informations section (Table S1).

Before and after sampling, filters were 48h-equilibrated (35% RH, T ambient) and weighted with a microbalance (1µg precision, model M5P-000V001 Sartorius, Germany) to measure the particle concentration (µg m^{-3}). All sampled filters were then preserved in the dark at –20°C (to avoid photodegradation and evaporation processes) for the chemical analyses.

PM sample preparation and extraction. Filters were prepared in order to perform different chemical analysis on the same samples. From each daily filter, 4 spots were cut: 1 spot was used for the determination of both n-alkanes and polycyclic aromatic hydrocarbons (PAHs) (results not reported) by GC-MS. In order to obtain enough PM mass for trace organic chemical analyses such as n-alkanes and PAHs, 3 daily filters were usually pooled together.

For n-alkane (and PAH) analysis, the PM filters were extracted in dichloromethane (CH_2Cl_2, purity ≥ 99.8%, Ultra Resi-Analyzed, J.T. Baker) in an ultrasonic bath (Sonica®, Soltec). PM samples were placed in an amber glass vial containing 2 ml of CH_2Cl_2 and ultrasonically extracted once for 20 minutes. The extract was then filtered through a PTFE syringe filter (cut 0.45 µm, Alltech) to remove insoluble particles. The extraction solvent was evaporated under a gentle stream of nitrogen (N_2, purity ≥ 99.9999%, Sapio) until dry. The residue was dissolved in 200 μl of isooctane (C_8H_8, purity ≥ 99.5%, for residue analysis, Fluka), transferred to a 1 ml amber glass vial with PTFE septa and kept at –20°C in the dark to prevent losses and photochemical reactions until the GC-MS analysis.

PM sample Gas Chromatograpy/Mass Spectrometry (GC/MS) analysis. The extracts were analysed by Gas Chromatography (GC) coupled with Mass Spectrometry (MS). An Agilent 6850 GC was used equipped with autosampler and a split/splitless injector. The separation was performed on a DB-XLB capillary column (length 60 m, i.d. 250 µm, film 0.25 µm; J&W Scientific). The injector was kept at 300°C and 2 μl of extract were injected in splitless mode. Helium (He; purity 99.999%, Sapio) was used as carrier gas with a constant flow of 1 ml/min. The n-alkane analysis was performed under the following temperature programme: (1) temperature ramp from 60° to 300°C at 6° C min^{-1}, (2) isothermal hold at 300°C for 20 min. The transfer line was kept at 305°C.

A quadrupole mass spectrometer (5973 Network Mass Selective Detector, Agilent Technologies) was used and operated at 70 eV in the electron ionization (EI) mode. The chromatograms were acquired in the SIM (Single Ion Monitoring) mode by monitoring 57, 71 and 85 m/z values during the whole chromatographic run.

Identification and quantification of n-alkanes. A standard mixture of n-alkanes (C_{14}-C_{32}) was prepared from single solid standards purchased from Alltech and diluted in isoctane (C_8H_8; purity $\geq$99.5%, for residue analysis, Fluka). This mixture was then diluted to obtain a concentration range of 40-0.02 $\mu g/ml$ to be used to compute the calibration curve for each n-alkane. The external standard method was used for n-alkane quantification based on the GC-MS signal obtained from the sum of three m/z values (57, 71 and 85). The calibration curves obtained for the C_{14}-C_{32} n-alkanes show good linearity with intercept values close to zero and regression coefficients R^2 higher than 0.995.

The method detection limit (X_{LOD}) was calculated from analysis of blank field filters (filters not used in PM sampling but submitted to the same manipulation as the samples used in field and laboratory experiments): X_{LOD} values were computed as the mean signal of all analysed blank field filters (n=7) plus three times the deviation standard. The detection limit ranged from 0.07 ng/m^3 for n-eicosane (C_{20}) to 0.08 ng/m^3 for n-dotriacontane (C_{32}) when concentration is reported to the average sampling volume of 41.4 m^3 (the pool of 3 different quarters of daily filters). The obtained detection limits were compatible with monitoring of n-alkane comcentrations in PM samples.

The n-alkanes were identified by matching the retention times of each peak in the sample chromatogram with those of a standard solution. Interfering coeluition problems were evaluated in the samples by comparing mass spectra of the samples with those of the standards as well as with those from the NIST mass spectra library (NIST MS Search r. 2.0). Problems of interference were only found for the n-C_{19} peak at m/z=57. Therefore, the C_{19} peak area was quantified in standard solutions and samples by using the sum of m/z=71 and m/z=85, excluding the m/z=57 contribution.

Computations on GC/MS signals. The algorithms used for the signal processing of the GC-MS data are written in the *MATLAB* ® (The MathWorks, Inc., R2007b) package. Computations were performed on a 1.53 GHz (256 RAM), AMD Athlon personal computer.

The first step of data handling consisted of a procedure to linearize the chromatographic signal to obtain constant retention increments between subsequent terms of the homologous series (eq. 4). It is a retention time alignment algorithm based on comparison vs. n-alkane standard mixture [26].

The Autocovariance Function was then numerically calculated from the linearizated chromatogram, according to eq. 1. A *MATLAB* algorithm was implemented, based on eqs. 5 and 10, to directly estimate the parameters n_{max}, CPI_{EACVF} and *series%* from the *EACVF* computed on a properly selected region of the chromatogram corresponding to the C_{14}-C_{32} n-alkanes.

Results and Discussion

Under the applied GC-MS analysis conditions, the n-alkanes ranging from C_{14} to C_{32} can be identified in the investigated PM samples. However, the lighter C_{14}-C_{19} terms were found at a low concentration level, lower than the detection limit for more than 50% of the samples. It must be noted that the first region of the chromatogram, where C_{14}-C_{19} n-alkanes elute, was quite disturbed due to coelution of other interfering compounds. Moreover, the lighter n-alkanes with $C \leq 19$ are generally considered too volatile to be accurately determined in PM samples as they incur evaporative losses during the sampling and analytical procedures [6,11,16].

For all the above reasons, the terms ranging from C_{20} to C_{32} were investigated in the present study as potential tracers for biogenic/antropogenic emissions: they were detected in all the analyzed PM samples, displaying a concentration level higher than the detection limit for most of the samples (>80%).

Spatial and seasonal variations of n-alkane content in PM samples. The concentrations of each C_{20}-C_{32} term were added together to obtain the total value $\Sigma C_{20} - C_{32}$ which was then used as an estimate of the n-alkane content in airborne particulate samples (reported in Table 1, 5th column). In addition to the concentration level (ng/m^3), a relative $\Sigma C_{20} - C_{32}\%$ value was also computed, as the ratio between the $\Sigma C_{20} - C_{32}$ values and the particle concentrations: it represents the partial contribution of n-alkanes to the PM total mass (Table 1, 6th column).

To investigate the variation in total n-alkane $\Sigma C_{20} - C_{32}$ concentrations for different sampling sites, seasons and particle sizes ($PM_{2.5}, PM_1, PM_{10}$), the mean values were computed (± mean SD, calculated as $SD/\sqrt{n}$) and reported in Table 2.

Lower concentrations were found in summer for the $PM_{2.5}$ samples collected in all the 3 sites: $\Sigma C_{20} - C_{32}$ values were 8.8 (± 0.9) ng/m^3 in MI, 8.9 (± 1.1) ng/m^3 in OB and 5.9 (± 0.4) ng/m^3 in ASC. In summer, the concentrations observed at the urban site were of an order of magnitude similar to those observed at clean rural and remote sites. In fact, in Northen Italy, summer is characterised by conditions of atmospheric instability which facilitate atmospheric transport, thus making the fine PM concentrations uniform throughout the entire region [28]. In this season, the high altitude remote site (ASC) is within the boundary layer and therefore affected by atmospheric transport from the plains.

The highest n-alkane concentrations were encounterd during fall and winter in the urban site (MI), ranging from 95.9 (± 4.7) ng/m^3 in PM_1 samples (fall) to 194.9 (± 20.5) ng/m^3 in PM_{10} samples (winter). This result is consistent with a restricted atmospheric transport of PM from source areas (like the city) to near ground (like rural area) due to the stable atmospheric conditions present in fall and winter that confine vertical distribution of pollutants to the first hundred meters of the atmosphere [29]. The partial contribution of n-alkanes to the PM total mass ($\Sigma C_{20} - C_{32}\%$) is low since it falls within the 0.035- 0.566% range (6th column in Table 1). The obtained results are consistent with those found for PM_{10} in urban sites in winter in other studies: e.g. 0.05-0.15% ($\Sigma C_{24} - C_{33}\%$) Vienna [30], 0.26% ($\Sigma C_{19} - C_{33}\%$) Taiwan [12]. Taiwan values are like the ones we measured in MI (0.27% for winter PM10 samples). It must be noted that, althougth the n-alkanes represent only minor costituents of PM, they contain very relevant information, helpful for input source characterization.

The $\Sigma C_{20} - C_{32}\%$ value has the same seasonal trend as $\Sigma C_{20} - C_{32}$ concentrations: minimum values were encountered in summer in all the 3 sites, with maximum in fall and winter. $\Sigma C_{20} - C_{32}\%$ values were 0.081 (± 0.09) % in MI, 0.113 (± 0.021) % in OB and 0.211 (± 0.056) in ASC for summer $PM_{2.5}$ samples: for the urban site (MI), fall (0.308 %) and winter (0.264 %) values were more than 3 times those found in summer. Such seasonality could be a result of the higher emissions from combustion, a major source of n-alkanes in the cold seasons [7,8,30].

Particle size variations of n-alkane content in PM samples. Since different cutpoint inlets were simultaneously used for collecting PM samples (2.5 μm and 1 μm in all the sampling sites and 10 μm cutpoint in MI), the obtained results make it possible to investigate n-alkane particle size distribution. The two PM dimensional fractions PM_1 and $PM_{2.5}$ of fine PMs showed quite similar n-alkane concentrations (Table 2, 7th and 9th columns). When the n-alkane concentrations were measured in PM_1 and $PM_{2.5}$ samples collected simultaneously (from the same site, on the same days), the $\Sigma C_{20} - C_{32}$ values found in PM_1 were on the average 68.5% (± 4.2%) of those in $PM_{2.5}$. That means that about 70% of the n-alkane concentration in $PM_{2.5}$ samples is really found in the finest submicrometric PM fraction (PM_1).

PM_{10} was also sampled and analysed in MI site (Table 2, 3rd and 4th columns). In this case, the ambient concentration of n-alkanes measured in PM_{10} was meanly 55.2% (± 5.3%) of that in simultaneously sampled $PM_{2.5}$. By sampling fine PM ($PM_{2.5}$), about 50% of total n-alkane concentration in PM_{10} is taken into account: the other half (50%) is encountered in the coarse fraction PM_{10}-$PM_{2.5}$ This is important to keep in mind when comparing the concentration of n-alkanes in atmospheric PM, referring to different PM dimensional fraction.

Compared to these results, the literature reports slightly higher abundances of *n*-alkanes in the fine PM. For example, Bi [13] assessed that more than 80% of the total concentrations of *n*-alkanes (C_{15} to C_{35}) were accumulated in particles <1.5 μm, both in urban and rural sites of Guangzhou (China). A distribution of the *n*-alkanes (C_{19} to C_{33}) in PM_1/PM_{10} is reported in the range of 0.66 to 0.88 for the urban site of Taiwan [12].

n-alkane distribution: carbon preference index (CPI). The CPI parameter was computed using the traditional procedure based on peak integration of the $C_{20}-C_{33}$ n-alkane GC-MS signal to describe their abundance distribution [10] (CPI_{trad}, 7th column in Table 1). Most of the analyzed samples show $CPI \cong 1$ values indicating strong contribution of emissions from urban ''winter'' sources, such as domestic heating (e.g., natural gas, oil, and wood combustion) generating a random distribution of odd/even terms of the series. On the other hand, the summer samples show higher CPI values (1.5-3.5) due to the higher contribution of the odd terms C_{27}, C_{29} and C_{31} originating from plant material which yield maximum emissions during the vegetative season [5, 30]. The means (± mean SD) of CPI values were computed for the three sampling sites at different seasons (Table 2, 5th, 10th columns). Mean summer CPI values for fine PM samples (PM_1 and $PM_{2.5}$) are 1.5 in MI and 1.2 in OB and ASC. A higher value was encountered for the PM_{10} sample, i.e. 2.5 (0.7) in MI. This would suggest that the main contribution to n-alkanes from plant material is found in the coarse PM fraction, and this is resonable as it is supposed a primary source of this is mechanical leaf abration (plant debris). In the literature, it is generally reported that the *n*-alkane of natural origin are predominantly found in the coarse PM, while those of anthropogenic origin tend to be found in the fine fraction [12,31].

Data handling of GC-MS signals using the $EACVF$ method. The chemometric method was applied to all chromatographic signals from GC-MS analysis of the PM samples. The aim was to test the method's ability to characterize the n-alkane contribution, in terms of CPI, and its applicability as an high-throughput method for analysis of the huge amounts of data from environmental monitoring. In comparison with the traditional procedure based on computation performed on integrated chormatographic peaks, the $EACVF$ method displays three fundamental advantages:

- it saves time and labor in data handling, thus increasing throughput and flexibility;
- it increases result reliability by deconvolving complex signals into its components;
- it reduces the subjectivity of human intervention, thus improving data quality.

These properties were investigated on a wide variety of samples in terms of sample particle size (PM_1, $PM_{2.5}$, PM_{10}), seasonality (winter vs. summer), and site location (urban vs. suburban).

The $EACVF$ was directly computed on the GC-MS signal (SIM signal at m/z values of 57, 71 and 85): the region 30-60 min was selected, since it contains the $C_{20}-C_{32}$ n-alkanes (Figure 1a, sample MI17). In comparison with the complex original GC signal, the $EACVF$ plot (Figure 1b) shows a simplified pattern characterized by a sequence of deterministic peaks located at $\Delta t=2.8$ min, i.e. the retention incrementen between subsequent terms of the the n-alkane series ($\Delta t=b$, eq.1) under the experimental GC conditions used. Such $EACVF$ peak is diagniostic, directly identifying the presence of n-alkanes and hence there is no need to compare them with the GC retention times of the reference standards ($C_{20}-C_{32}$).

The main information on n-alkane series are directly extracted from the values of the $EACVF$ computed at $\Delta t=kb$, for chracteristic k values.

The number n_{max} of n-alkanes present in the sample can be directly estimated from the $EACVF$ peaks at $\Delta t=kb$ for even k: for all the investigated GC-MS signals the n_{max} values were correctely estimated as $n_{max}=14$.

An $EACVF(2b)/EACVF(0)$ ratio was computed to estimate the relative contribution of the homologous series terms to the total amount of alkanes ($series\%$, 10th column in Table 1).

The abundance distribution of the odd/even terms, can be quantified by computing CPI_{EACVF} values directly from $EACVF$ using the values at $\Delta t=b$ and $\Delta t=2b$ (CPI_{EACVF}, 8th column in Table 1).

Many investigated signals display a high contribution of the UCM hump (Fig. 2a) because, in an effort to obtain a fast analytical procedure for n-alkane determination, the samples were obtained by a simple solvent extraction, without any extract purification. These conditions may yield an ambigous n-alkane characterization as a consequence of the coelution yielding of complex superimposed signals: this is particularly true for samples containing low abundance n-alkanes as expressed by $series\%$ values lower than 8. These chromatograms were handled with the complete procedure for deconvolving the UCM contribution from $EACVF_{tot}$ and computing CPI_{EAVF} from $EACVF_{res}$ (CPI_{EACVF} marked by a star in 8th column in Table 1).

Reliability of the CPI_{EACVF} data. The accuracy of the obtained results was checked by comparing the CPI_{EAVF} with CPI_{trad} values obtained using the two procedures (CPI_{trad} vs. CPI_{EACVF}, 7th, 8th columns in Table 1) and estimating the relative estimation error ($\varepsilon\%$, 9th column in Table 1).

In general, a good agreement was achieved between the two procedures: the relative error $\varepsilon\%$ was lower than 15% for 70 of the 76 investigated samples and lower than 5% for 22 of them. The exceptions are 6 PM samples for which different CPI_{trad} and CPI_{EACVF} values were estimated ($\varepsilon\% \geq$ 15%). They correspond to the samples containing the lowest n-alkane abundance ($series\% \leq 4$) and which generate complex GC signals showing coeluting components and superimposed UCM band: this makes it very difficult to integrate the n-alkane peaks and to deconvolute the $EACVF$ plot to prevent an unbiased estimation of the CPI_{trad} and CPI_{EACVF} values.

It must be noted that the whole procedure is reliable in estimating accurate parameters, since it also includes retention time rescaling and UCM component subtraction: all these results are directly obtained from the whole chromatographic signal by an algorithm requiring just a few minutes of PC computation time.

Characterization of the PM samples by $EACVF_{tot}$ method. The parameters CPI_{EACVF} and $series\%$ derived from $EACVF$ for each PM sample were investigated to extract information on source contributions to n-alkanes (plots reported in Figs. 3a and 3b). The highest PM classification is based on seasonal differences and is nearly independent of sampling site and particle size. Such a characterization is due to large variations in the relative abundance of n-alkanes, as represented by $series\%$ values ranging from 1% to 48%: the lowest values ($series\% =$ 1-14%) were found in summer and the highest ($series\% =$24-48%) in winter, most likely the result of higher emissions from urban ‘‘winter’’ sources [4,7,8]. In addition to this classification, a considerable seasonal variation in CPI_{EACVF} values can be identified: CPI_{EACVF} values close to 1 ($CPI_{EACVF} = 0.8$-1.2) were found in the cold season for all the sampling sites, revealing a strong contribution from anthropogenic emissions. A greater spread in CPI_{EACVF} values ($CPI_{EACVF} = 0.9$-3) was observed in the warm season showing the influence of the contribution from biogenic sources in all the sampling sites.

This classification is sharper if the same PM dimensional fraction is considered: data for $PM_{2.5}$ are reported in Fig. 3b. The values obtained show strong seasonal differentiation between samples collected during the summer — and which contain lower n-alkane concentrations ($series\% =$ 1-11%) — and those collected in the fall-winter — containing a significantly higher n-alkane contribution

($series\% = 18$-43%). In addition to this classification, a considerable seasonal variation in CPI_{EACVF} values can be identified for the $PM_{2.5}$ samples.

In conclusion, the presented results provide experimental evidence that the $EACVF$ procedure is robust, able to extract reliable information on the n-alkane distribution from direct handling of complex GC-MS chromatograms, such as SIM and TIC signals, also containing superimposed UCM hump: it can increase GC-MS analysis throughput and flexibility without sacrificing data quality or reliability of the results.

Therefore, it can be proposed as a reliable alternative to the cumbersome, and time-consuming, procedure based on chromatogram integration thus offering simple, quick characterization of n-alkane distribution patterns. This property is especially helful for characterizing the distribution patterns of homologous series as chemical tracers in organics input sources to be used whenever attempting to identify the origin of an aerosol for the purpose of pollution control or abatement.

Literature Cited

[1] Mazurek M.A., 2002. *Molecular Identification of Organic Compounds in Atmospheric Complex Mixtures and Relationship to Atmospheric Chemistry and Sources.* **Environmental Health Perspectives** 110, 995-1003.

[2] Alves C., Oliveira T, Pio C., Silvestre A.J.D., Fialho P., Barata F., Legrand M., 2007. *Characterization of carbonaceous aerosols from the Azorean Island of Terceira.* **Atmospheric Environment** 41, 1359–1373.

[3] Fu P., Kawamura K., Barrie L., 2009. *Photochemical and Other Sources of Organic Compounds in the Canadian High Arctic Aerosol Pollution during Winter-Spring* **Environmental Science & Technology** 43, 286–292.

[4] Wang G., Kawamura K., Xie M., Hu, S., Cao J., An Z., Waston J.G., Chow J.C., 2009. *Organic Molecular Compositions and Size Distributions of Chinese Summer and Autumn Aerosols from Nanjing: Characteristic Haze Event Caused by Wheat Straw Burning.* **Environmental Science & Technology** 43, 6493–6499.

[5] Simoneit B.R.T., 1984. *Organic matter of the troposphere III. Characterization and sources of petroleum and pyrogenic residues in aerosols over the Western United States.* **Atmospheric Environment** 18, 51-67.

[6] Mandalakis M., Tsapakis M., Tsoga A., Stephanou E.G., 2002. *Gas-particle concentrations and distribution of aliphatic hydrocarbons, PAHs, PCBs and PCDD/Fs in the atmosphere of Athenes (Greece).* **Atmospheric Environment** 36, 4023-4035.

[7] Li M., McDow S.R., Tollerud D.J., Mazurek M.A., 2006. *Seasonal abundance of organic molecular markers in urban particulate matter from Philadelphia, PA.* **Atmospheric Environment** 40, 2260–2273.

[8] Ladji R., Yassaa R., Balducci C., Cecinato A., Meklati B.Y., 2009. *Annual variation of particulate organic compounds in PM10 in the urban atmosphere of Algiers.* **Atmospheric Research** 92, 258–269.

[9] Cranwell, P. A., 1981. *Diagenesis of free and bound lipids in terrestrial detritus deposited in a lacustrine sediment.* **Organic Geochemistry** 3(3), 79-89.

[10] Bray E.E., Evans E.D., 1961. *Distribution of n-paraffins as a clue to recognition of source bed.* **Geochimica Cosmochimica Acta** 22, 2-15.

[11] Bi X., Sheng G., Peng P., Chen Y., Zhang Z., Fu J., 2003. *Distribution of particulate- and vapour-phase n-alkanes and polycyclic aromatic hydrocarbons in urban atmosphere of Guangzhou, China.* **Atmospheric Environment** 37, 289-298.

[12] Lin J.J., Lee L.-C., 2004. *Characterization of n-alkanes in urban submicron aerosol particles* (PM_1)*.* **Atmospheric Environment** 38, 2983-2991.

[13] Bi X., Sheng G., Peng P., Chen Y., Fu J., 2005. *Size distribution of n-alkanes and polycyclic aromatic hydrocarbons (PAHs) in urban and rural atmospheres of Guangzhou, China.* **Atmospheric Environment** 39, 477-487.

[14] Wang G., Huang L., Zhao X., Niu H., Dai Z., 2006. *Aliphatic and polycyclic aromatic hydrocarbons of atmospheric aerosols in five locations of Nanjing urban area, China.* **Atmospheric Research** 81, 54– 66.

[15] Cheng Y., Li S.-M, Leithead A., Brook J.R., 2006. *Spatial and diurnal distributions of n-alkanes and n-alkan-2-ones on PM2.5 aerosols in the Lower Fraser Valley, Canada.* **Atmospheric Environment** 40, 2706–2720.

[16] Cincinelli A., Del Bubba M., Martinelli T., Gambaro A., Lepri L., 2007. *Gas-particle concentration and distribution of n-alkanes and polyciclic aromatic hydrocarbons in the atmosphere of Prato (Italy).* **Chemosphere** 68, 472-478.

[17] Pietrogrande M.C., Mercuriali M., Pasti L., Dondi F., 2009. *Data handling of complex GC–MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source identification.* **Analyst** 134, 671-680.

[18] Christensen J.H., Mortensen J., Asger B.H., Andersen O., 2005. *Chromatographic preprocessing of GC–MS data for analysis of complex chemical mixtures.* **Journal of Chromatography A** 1062, 113–123.

[19] Amigo J.M., Skov T., Coello J., Maspoch S., Bro R., 2008. *Solving GC-MS problems with PARAFAC2.* **Trends in Analytical Chemistry** 27(8), 714-725

[20] Xu L., Tang L.-J., Cai C.-B., Wu H.-L., Shen G.-L., Yu R.-Q., Jiang J.-H., 2008. *Chemometric methods for evaluation of chromatographic separation quality from two-way data - a review.* **Analytica Chimica Acta** 613, 121–134.

[21] Shackmana J.G., Watson C.J., Kennedy R.T., 2004. *High-throughput automated post-processing of separation data.* **Journal of Chromatography A** 1040, 273–282.

[22] Felinger A., Pietrogrande M.C., 2001. *Decoding Complex Multicomponent Chromatograms.* **Analytical Chemistry** 73, 618A-622A.

[23] Marchetti N., Felinger A., Pasti L., Pietrogrande M.C., Dondi F., 2004. *Decoding Two-Dimensional Complex Multicomponent Separations by Autocovariance Function.* **Analytical Chemistry** 76, 3055-3068.

[24] Pietrogrande M.C., Zampolli M.G., Dondi F., Szopa C., Sternberg R., Buch A., Raulin J., 2005. *In situ analysis of the Martian soil by gas chromatography: Decoding of complex chromatograms of organic molecules of exobiological interest.* **Journal of Chromatography A** 1071, 255–261.

[25] Pietrogrande M.C., Zampolli M.G., Dondi F., 2006. *Identification and quantification of homologous series of compound in complex mixtures: Autocovariance study of GC/MS chromatograms.* **Analytical Chemistry** 78, 2576–2592.

[26] Pietrogrande M.C., Mercuriali M., Pasti L., 2007. *Signal processing of GC–MS data of complex environmental samples:Characterization of homologous series.* **Analytica Chimica Acta** 594, 128-138.

[27] Pietrogrande M.C., Mercuriali M., Bacco D., 2008. *Data handling of complex GC-MS signals to characterize homologous series as organic source tracers in atmospheric aerosols.* **Air Pollution XVI**, eds. C.A. Brebbia, J.W.S. Longhust, WITPress, 335-343.

[28] Vecchi R., Marcazzan G., Valli G., Ceriani M., Antoniazzi C., 2004. *The role of atmospheric dispersion in the seasonal variation of PM1 and PM2.5 concentration and composition in the urban area of Milan (Italy).* **Atmospheric Environment** 38, 4437-4446.

[29] Ferrero L., Petraccone S., Perrone M.G., Sangiorgi G., Lo Porto C., Lazzati Z. and Ferrini B., 2007. *Vertical profiles of parti culate matter over Milan during winter 2005/2006.* **Fresenius Environmental Bulletin**, 16, 697-700.

[31] Karanasiou A.A., Sitaras I.E., Siskos P.A., Eleftheriadis K., 2007. *Size distribution and sources of trace metals and n-alkanes in the Athens urban aerosol during summer*. **Atmospheric Environment** 41, 2368-2381.

[30] Kotianová P., Puxbaum H., Bauer H., Caseiro A., Marr I.L., Čík G., 2008. *Temporal patterns of n-alkanes at traffic exposed and suburban sites in Vienna*. **Atmospheric Environment** 42, 2993-3005.

Table 1. Properties of 22 investigated PM samples collected in the urban site of Milan (MI), categorized according to the particle size ($PM_{2.5}$, PM_{1}, PM_{10}, 2nd column). The reported parameters are: PM concentration (ng/m^3, 4th column), n-alkane concentration (ng/m^3, 5th column) and relative value to the PM amount (6th column), CPI values computed according to the traditional procedure (CPI_{trad}, 7th column) and the $EACVF$ method (CPI_{EACVF}, 8th column), $\varepsilon\%$ (9th column) to estimate the relative estimation error of CPI_{EAVF} vs. CPI_{trad} values and $series\%$ (10th column) which is the relative contribution of the homologous series terms to the total amount of alkanes.

CPI_{EACVF} values marked by star in the 8th column indicate CPI_{EACVF} values computed from $EACVF_{res}$ using the complete procedure.

Sample	PM_x	*Season*	*PM* (ng/m^3)	ΣC_{20}-C_{32} (ng/m^3)	ΣC_{20}-C_{32}/ $PM\,\%$	CPI_{trad}	CPI_{EACVF}	$\varepsilon\%$	$series\%$
MI-1	$PM_{2.5}$	SUMMER	13.2	4.58	0.035	1.48	1.70*	14.5	4.10
MI-2	$PM_{2.5}$	SUMMER	5.10	3.86	0.076	1.47	1.02*	30.5	2.76
MI-3	$PM_{2.5}$	SUMMER	4.50	6.22	0.138	1.27	1.36	7.10	4.60
MI-4	$PM_{2.5}$	SUMMER	7.60	8.11	0.107	1.38	1.00*	27.6	2.21
MI-5	$PM_{2.5}$	SUMMER	10.7	7.76	0.073	1.61	1.87*	15.9	4.04
MI-6	$PM_{2.5}$	SUMMER	14.9	7.13	0.048	1.43	1.60*	11.7	3.29
MI-7	$PM_{2.5}$	SUMMER	15.3	8.60	0.056	1.56	1.50	3.95	6.93
MI-8	$PM_{2.5}$	SUMMER	15.3	9.46	0.062	1.48	1.38	6.67	6.61
MI-9	$PM_{2.5}$	SUMMER	12.4	7.73	0.062	1.20	1.05	12.5	8.85
MI-10	$PM_{2.5}$	SUMMER	8.80	9.63	0.109	1.64	1.48*	9.75	7.00
MI-11	$PM_{2.5}$	SUMMER	11.8	16.0	0.136	1.69	1.45*	14.2	8.70
MI-12	$PM_{2.5}$	SUMMER	14.1	10.5	0.074	1.89	1.71*	9.60	8.69
MI-13	$PM_{2.5}$	SUMMER	18.3	14.2	0.078	1.88	1.61*	14.1	11.1
MI-14	PM_{1}	SUMMER	8.30	8.89	0.107	1.31	1.17*	10.7	9.01
MI-15	PM_{1}	SUMMER	15.3	11.3	0.074	1.77	1.71	3.25	6.10
MI-16	PM_{1}	SUMMER	12.8	9.40	0.073	1.59	1.62	1.64	6.29
MI-17	PM_{1}	SUMMER	5.40	28.2	0.522	1.23	1.13	8.10	46.2
MI-18	PM_{1}	SUMMER	8.10	5.78	0.071	1.76	1.53*	13.0	6.02
MI-19	PM_{1}	SUMMER	9.50	5.41	0.057	1.68	1.54*	8.20	8.02
MI-20	PM_{10}	SUMMER	35.4	23.0	0.065	3.01	2.70*	10.2	32.7
MI-21	PM_{10}	SUMMER	19.2	16.9	0.088	3.38	2.98*	11.9	23.1
MI-22	PM_{10}	SUMMER	29.7	13.5	0.046	1.18	1.14	3.31	13.9

Table 2. Seasonal variation of PM ($\mu g/m^3$), n-alkane concentrations ($\Sigma C_{20} - C_{32}$: ng/m^3) and CPI_{trad} values for different particle sizes ($PM_{2.5}, PM_1, PM_{10}$) in urban (MI), rural (OB) and high altitude remote (ASC) sites. The mean values are reported (± mean SD, calculated as $SD/\sqrt{n}$).

		PM_{10} *Samples*			$PM_{2.5}$ *Samples*		PM_1 *Samples*		$PM_{2.5}$ PM_1 *Samples*
		PM *($\mu g/m^3$)*	*ΣC_{20}-C_{32}* *(ng/m^3)*	CPI_{trad}	*PM* *($\mu g/m^3$)*	*ΣC_{20}-C_{32}* *(ng/m^3)*	*PM* *($\mu g/m^3$)*	*ΣC_{20}-C_{32}* *(ng/m^3)*	CPI_{trad}
MI	SUMMER	31.3 (± 1.3)	17.8 (± 2.8)	2.5 (± 0.7)	16.8 (± 0.6)	8.8 (± 0.9)	10.4 (± 0.6)	8.2 (± 1.0)	1.5 (± 0.0)
	FALL	-	-	-	46.3 (± 3.6)	124.2 (± 8.8)	32.9 (± 2.4)	95.9 (± 4.7)	1.2 (± 0.0)
	WINTER	73.4 (± 6.7)	194.9 (± 20.5)	1.1 (± 0.1)	54.5 (± 3.5)	124.2 (± 14.0)	34.2 (± 3.6)	116.2 (± 36.7)	1.1 (± 0.0)
OB	SUMMER	-	-	-	15.5 (± 0.7)	8.9 (± 1.1)	10.8 (± 0.5)	8.5 (± 0.5)	1.2 (± 0.1)
	FALL	-	-	-	31.8 (± 4.0)	38.6 (± 2.7)	17.3 (± 2.3)	26.1 (± 5.9)	1.5 (± 0.2)
	WINTER	-	-	-	32.4 (± 2.3)	62.4 (± 13.9)	19.2 (± 1.6)	34.0 (± 6.8)	1.0 (± 0.0)
ASC	SUMMER	-	-	-	5.9 (± 1.1)	5.9 (± 0.4)	3.6 (± 0.4)	-	1.2 (± 0.1)
	WINTER	-	-	-	2.9 (± 0.3)	7.5 (± 1.4)	2.4 (± 0.4)	5.0 (± 1.3)	0.8 (± 0.0)

Figure captions

Figure 1: GC-MS signal of a PM sample with high n-alkane content (sample MI17, $series\%$=6).

Fig. 1a: SIM signal of the GC-MS chromatogram monitored at $m/z = 57+71+85$. The arrow in the peak at $\Delta t = 2b$ indicates the $EACVF_{res}(2k)$ value as the basis for estimating the $series\%$.

Fig. 1b: $EACVF$ plot computed on the chromatogram: lower solid line.

The arrows identify the peaks at $\Delta t = b = 1.80 min$ and $\Delta t = 2b = 3.60 min$ diagnostic in revealing the presence of the series terms.

Figure 2: GC-MS signal of a PM sample with low n-alkane content (sample MI20, $series\%$=18).

Fig. 2a: SIM signal of the GC-MS chromatogram monitored at $m/z = 57+71+85$;

Fig. 2b: $EACVF$ plot computed on the chromatogram: upper solid line; $EACVF_{UCM}$ plot computed on the unresolved component (UCM): dashed line; $EACVF_{res}$ plot computed on the resolved component of the chromatogram: lower bold line.

The arrows in the peaks at $\Delta t = 2b$ indicate the $EACVF_{res}(2k)$ value as the basis for estimating the $series\%$ value.

Figure 3: Characterization and classification of the different PM samples based on the parmeters computed by $EACVF$: relative abundance of the n-alkane series ($series\%$) and CPI_{EACVF} values.

Fig. 3a: data of all the PM samples investigated (76 samples);

Fig. 3b: data of PM samples having the same dimensional fraction (46 $PM_{2.5}$ samples).

Figure 1a

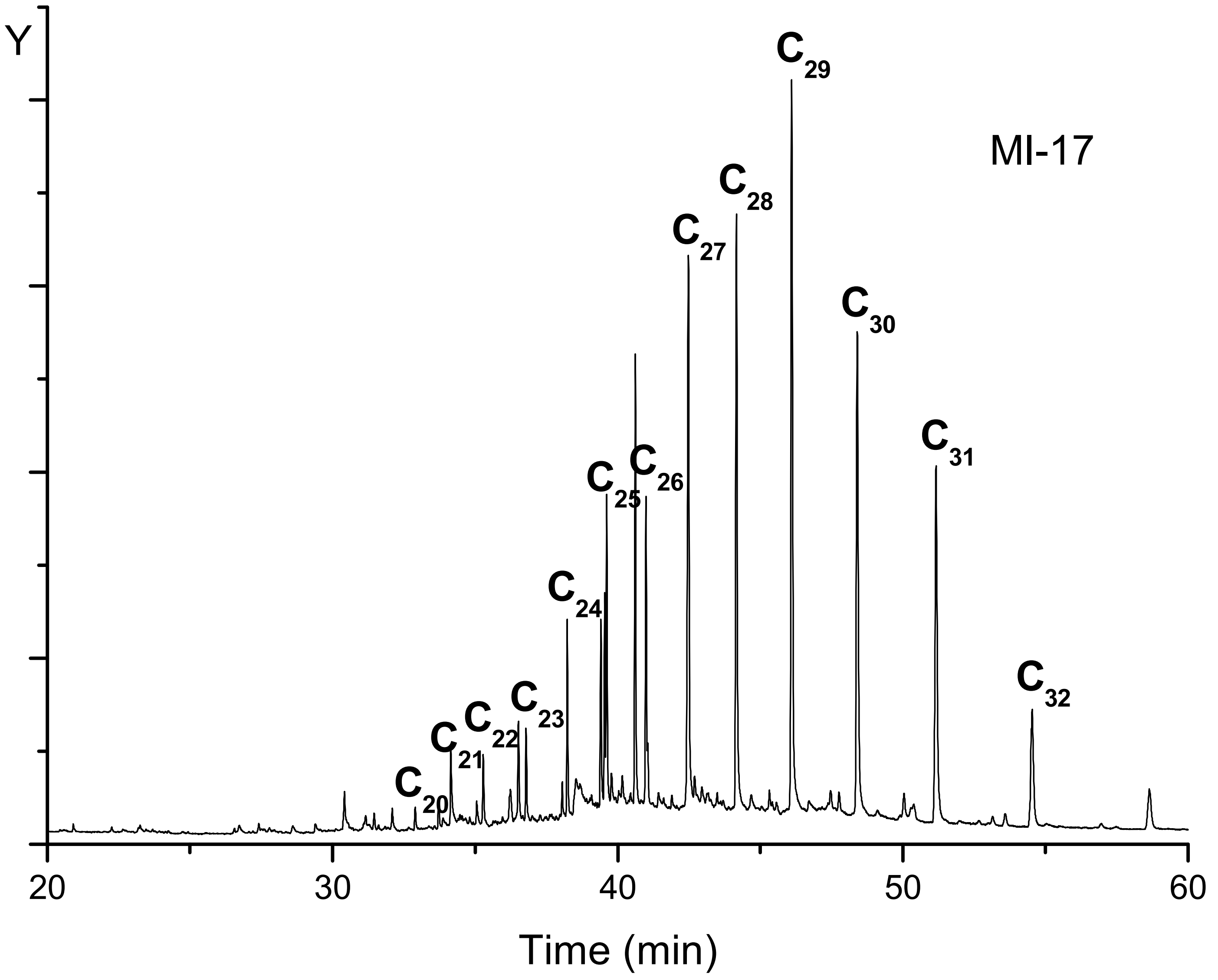

Y
MI-17
C20
C21
C22
C23
C24
C25
C26
C27
C28
C29
C30
C31
C32
20
30
40
50
60
Time (min)

Figure 1b

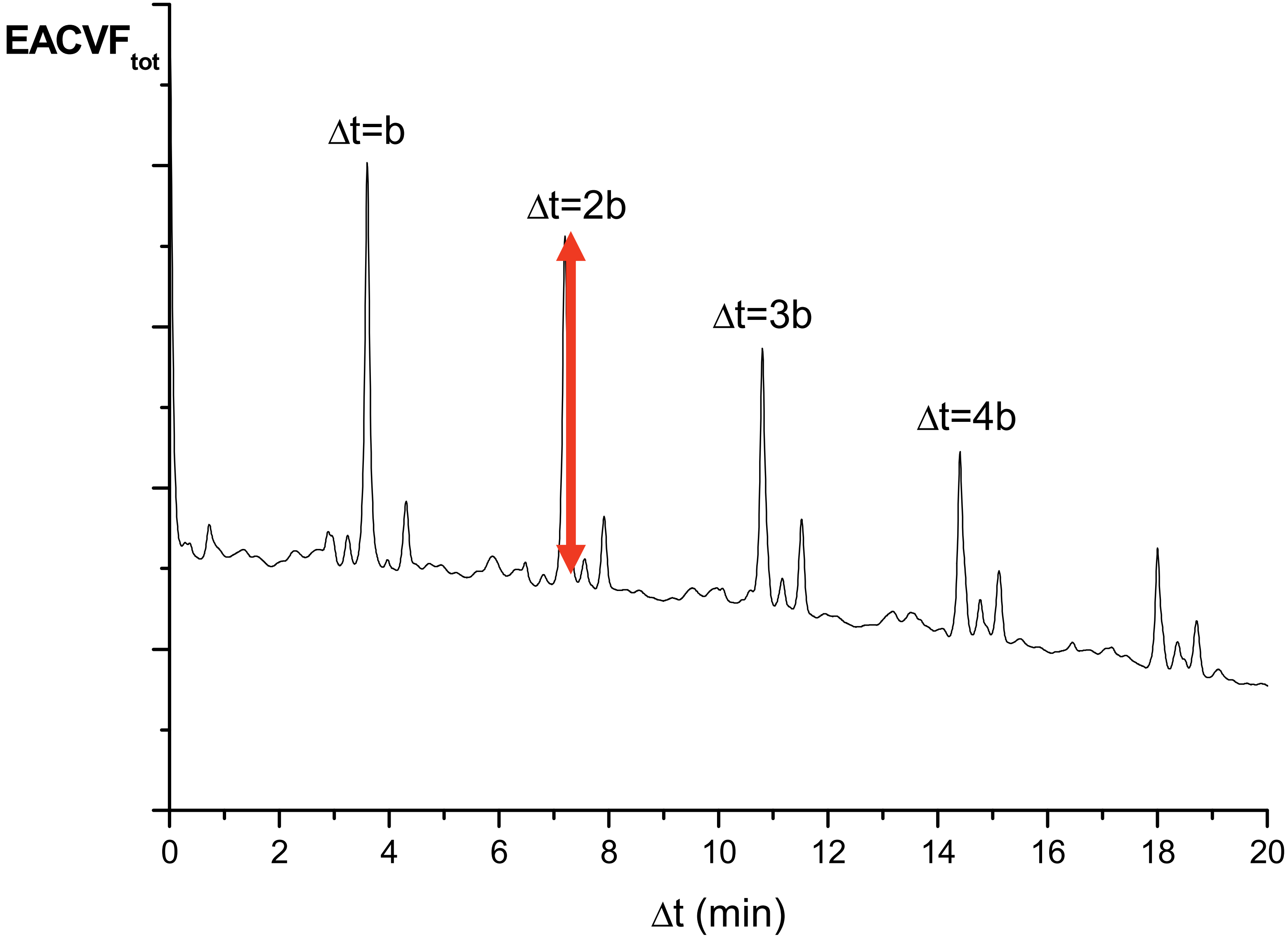


Figure 2a

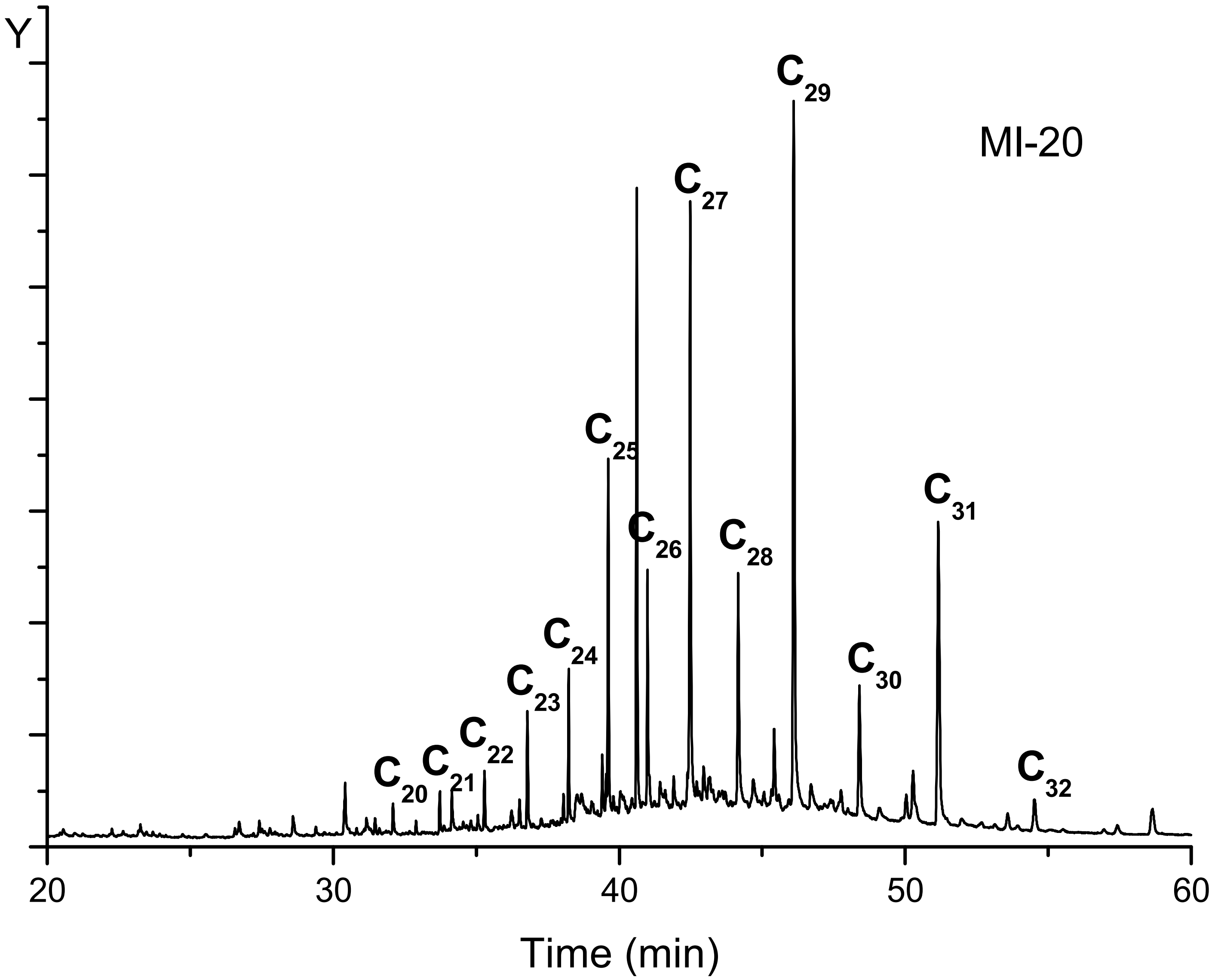

Y
MI-20
C_{20}
C_{21}
C_{22}
C_{23}
C_{24}
C_{25}
C_{26}
C_{27}
C_{28}
C_{29}
C_{30}
C_{31}
C_{32}
20
30
40
50
60
Time (min)

Figure 2b

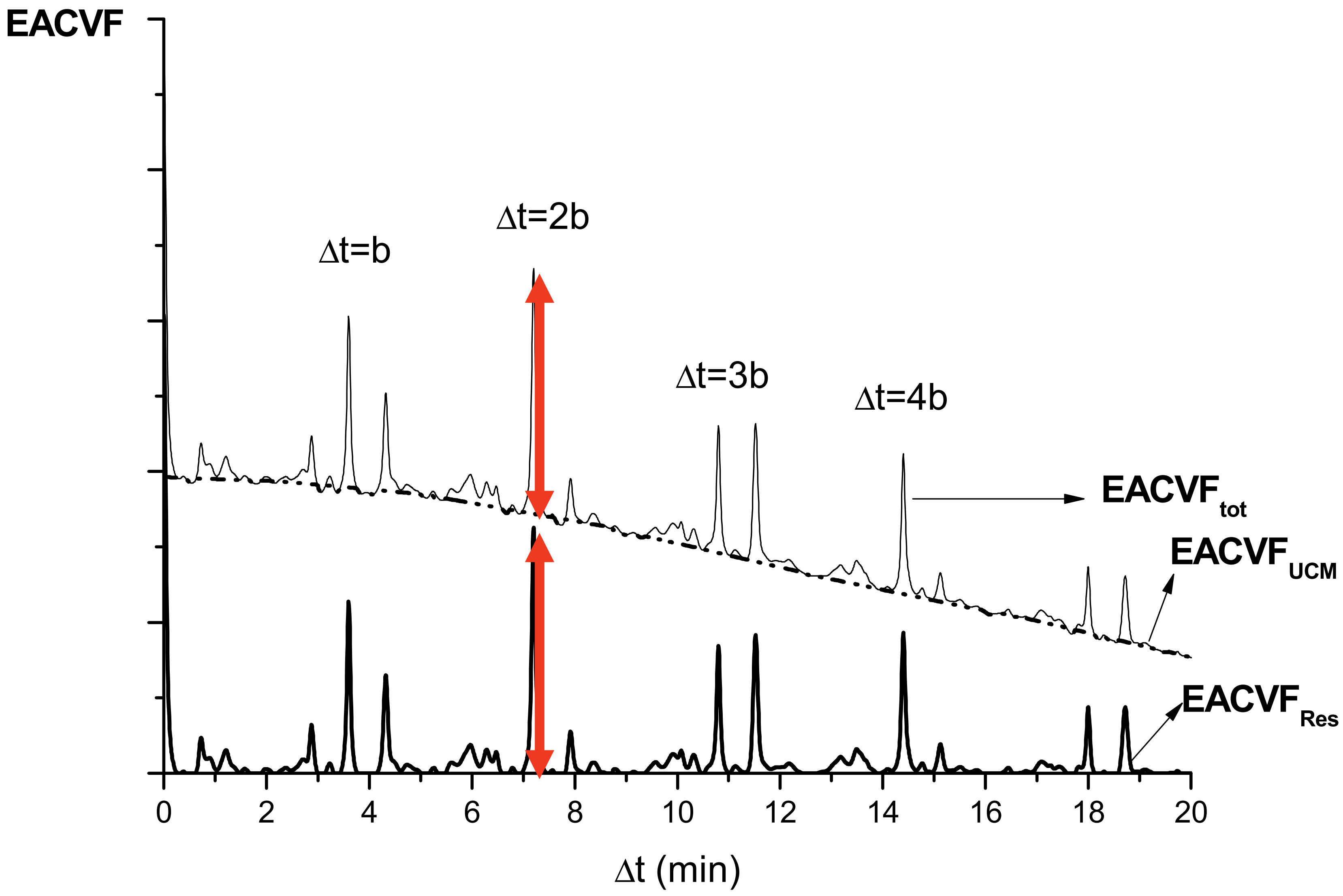

EACVF
Δt=b
Δt=2b
Δt=3b
Δt=4b
EACVF$_{tot}$
EACVF$_{UCM}$
EACVF$_{Res}$
0
2
4
6
8
10
12
14
16
18
20
Δt (min)

Figure 3a

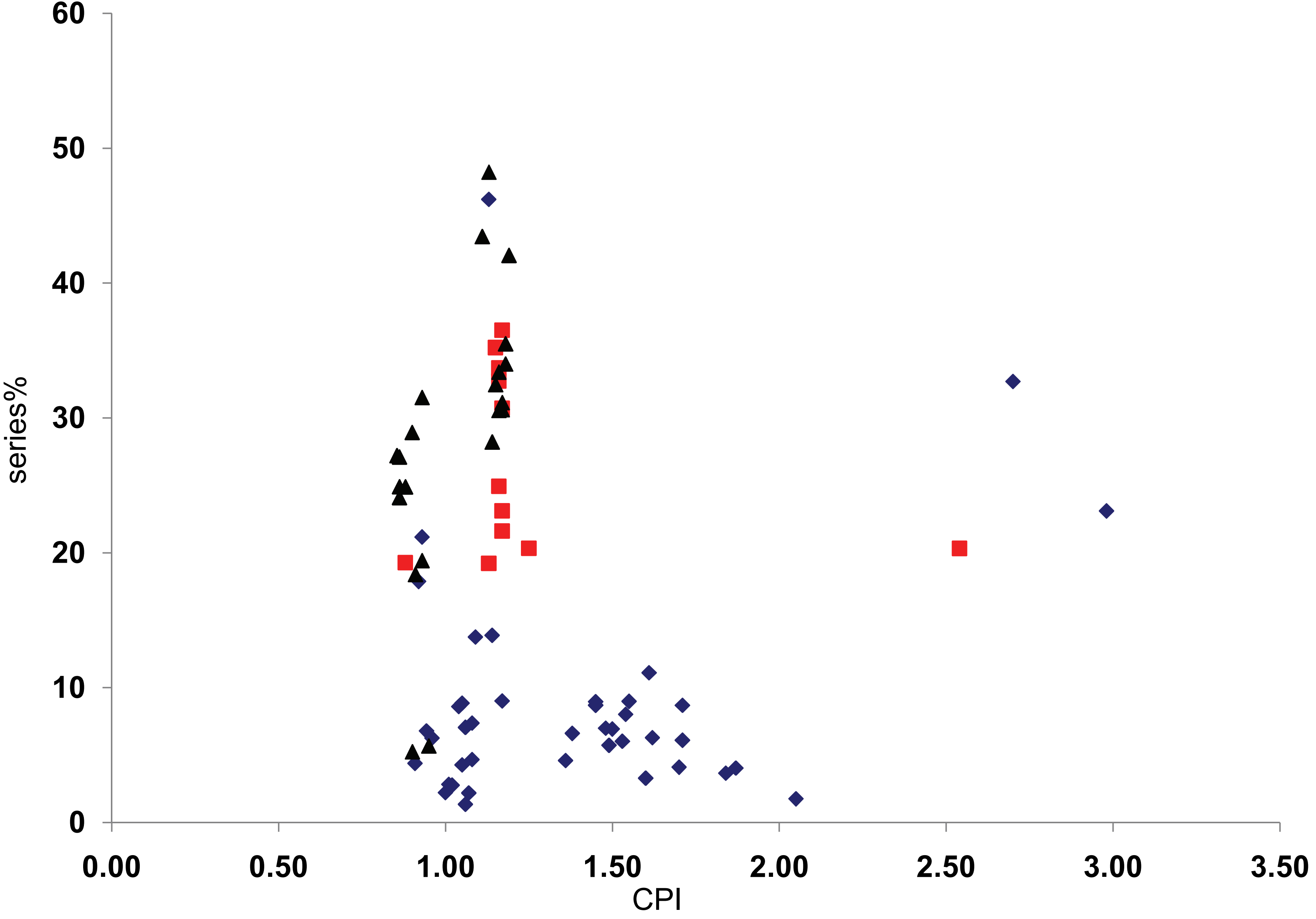
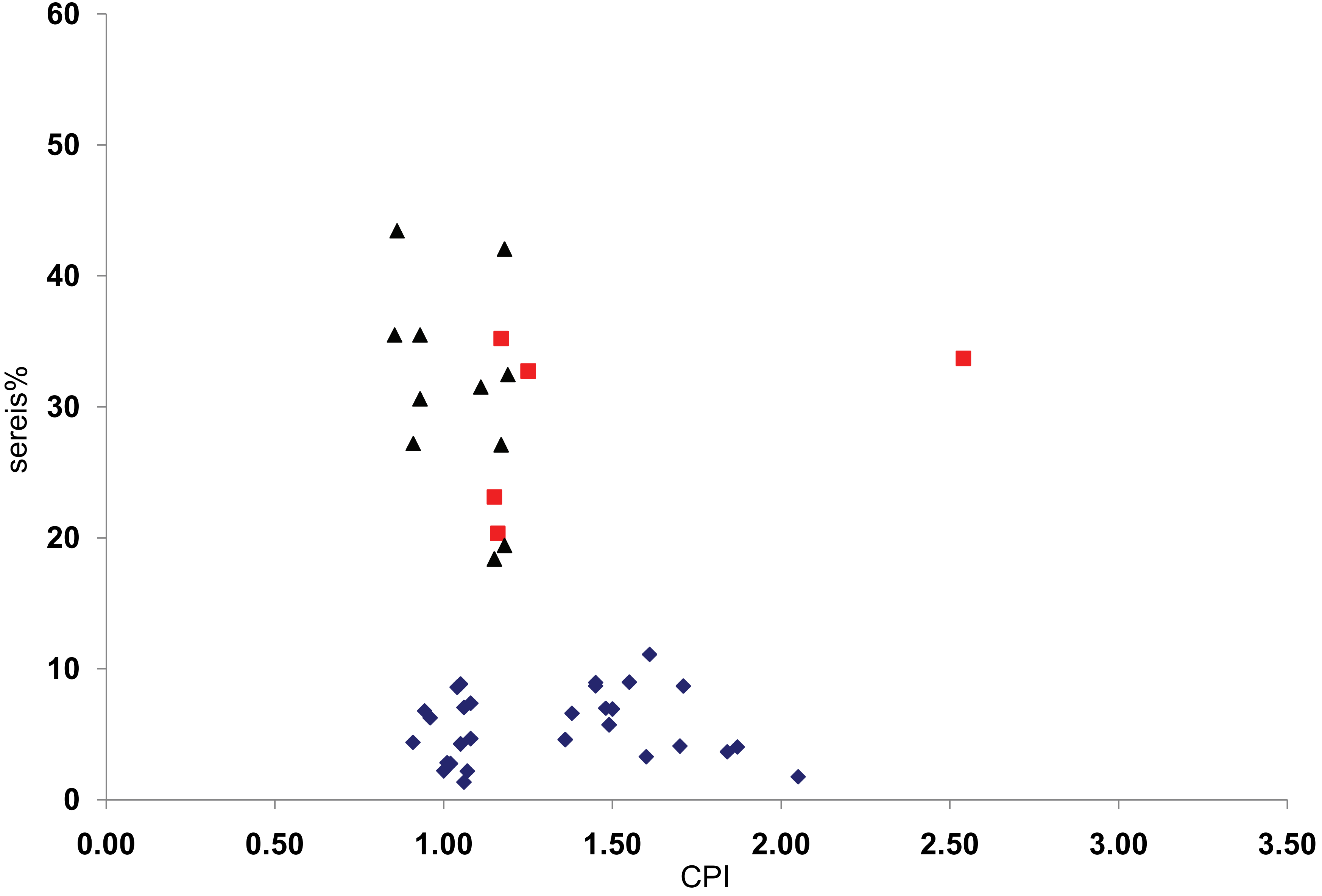


Figure 3b



Supporting Information

Supporting Information contains the following data:

Table S1: is the extendend form of the Table 1 in the main text containing the complete data set of all the 76 analyzed PM samples.
Data treatment: is a detailed description of the mathematical equations used in the developed algoritm for computing the properties of the homologous series directly from the *EACVF* values.

Sample	PM_x	*Season*	*PM* (ng/m^3)	ΣC_{20}-C_{32} (ng/m^3)	ΣC_{20}-C_{32}/PM %	CPI_{trad}	CPI_{EACVF}	*ε%*	*series%*
MI-1	$PM_{2.5}$	SUMMER	13.2	4.58	0.035	1.48	1.70*	14.5	4.10
MI-2	$PM_{2.5}$	SUMMER	5.10	3.86	0.076	1.47	1.02*	30.5	2.76
MI-3	$PM_{2.5}$	SUMMER	4.50	6.22	0.138	1.27	1.36	7.10	4.60
MI-4	$PM_{2.5}$	SUMMER	7.60	8.11	0.107	1.38	1.00*	27.6	2.21
MI-5	$PM_{2.5}$	SUMMER	10.7	7.76	0.073	1.61	1.87*	15.9	4.04
MI-6	$PM_{2.5}$	SUMMER	14.9	7.13	0.048	1.43	1.60*	11.7	3.29
MI-7	$PM_{2.5}$	SUMMER	15.3	8.60	0.056	1.56	1.50	3.95	6.93
MI-8	$PM_{2.5}$	SUMMER	15.3	9.46	0.062	1.48	1.38	6.67	6.61
MI-9	$PM_{2.5}$	SUMMER	12.4	7.73	0.062	1.20	1.05	12.5	8.85
MI-10	$PM_{2.5}$	SUMMER	8.80	9.63	0.109	1.64	1.48*	9.75	7.00
MI-11	$PM_{2.5}$	SUMMER	11.8	16.0	0.136	1.69	1.45*	14.2	8.70
MI-12	$PM_{2.5}$	SUMMER	14.1	10.5	0.074	1.89	1.71*	9.60	8.69
MI-13	$PM_{2.5}$	SUMMER	18.3	14.2	0.078	1.88	1.61*	14.1	11.1
MI-14	PM_1	SUMMER	8.30	8.89	0.107	1.31	1.17*	10.7	9.01
MI-15	PM_1	SUMMER	15.3	11.3	0.074	1.77	1.71	3.25	6.10
MI-16	PM_1	SUMMER	12.8	9.40	0.073	1.59	1.62	1.64	6.29
MI-17	PM_1	SUMMER	5.40	28.2	0.522	1.23	1.13	8.10	46.2
MI-18	PM_1	SUMMER	8.10	5.78	0.071	1.76	1.53*	13.0	6.02
MI-19	PM_1	SUMMER	9.50	5.41	0.057	1.68	1.54*	8.20	8.02
MI-20	PM_{10}	SUMMER	35.4	23.0	0.065	3.01	2.70*	10.2	32.7
MI-21	PM_{10}	SUMMER	19.2	16.9	0.088	3.38	2.98*	11.9	23.1
MI-22	PM_{10}	SUMMER	29.7	13.5	0.046	1.18	1.14	3.31	13.9
MI-23	$PM_{2.5}$	FALL	48.2	112	0.232	1.35	1.16	13.8	20.3
MI-24	$PM_{2.5}$	FALL	36.8	120	0.326	1.22	1.16	5.00	20.3
MI-25	$PM_{2.5}$	FALL	39.3	141	0.359	1.16	1.15	0.95	23.1
MI-26	PM_1	FALL	43.1	104	0.242	1.32	1.17	11.3	19.2
MI-27	PM_1	FALL	27.8	95.6	0.344	1.18	1.17	0.89	19.3
MI-28	PM_1	FALL	26.9	87.8	0.326	1.13	1.17	3.33	24.9
MI-29	PM_{10}	WINTER	93.3	198	0.212	1.06	1.17	10.2	31.1
MI-30	PM_{10}	WINTER	39.7	158	0.398	0.98	0.86	12.1	24.1
MI-31	PM_{10}	WINTER	118	229	0.194	1.17	1.16	1.00	30.6
MI-32	$PM_{2.5}$	WINTER	64.7	92.6	0.143	1.28	1.11	13.1	43.5
MI-33	$PM_{2.5}$	WINTER	33.7	85.1	0.253	1.15	1.15	0.41	32.5
MI-34	$PM_{2.5}$	WINTER	122	446	0.366	1.03	1.18*	15.0	35.5
MI-35	$PM_{2.5}$	WINTER	56.8	163	0.286	0.99	0.85	13.6	27.2
MI-36	$PM_{2.5}$	WINTER	35.4	132	0.372	1.17	1.17	0.15	30.6
MI-37	PM_1	WINTER	59.0	189	0.321	1.19	1.13	4.99	48.2
MI-38	PM_1	WINTER	34.0	72.7	0.214	1.11	1.14	2.41	28.2
OB-1	$PM_{2.5}$	SUMMER	13.1	8.39	0.064	1.37	1.55	13.4	8.99
OB-2	$PM_{2.5}$	SUMMER	7.60	7.99	0.105	1.34	1.49*	11.0	5.73
OB-3	$PM_{2.5}$	SUMMER	2.30	6.06	0.263	0.95	0.96	1.34	6.26
OB-4	$PM_{2.5}$	SUMMER	4.40	4.28	0.097	1.10	1.01*	8.09	2.83

OB-5	$PM_{2.5}$	SUMMER	9.00	10.0	0.111	1.27	1.45	14.1	8.96
OB-6	$PM_{2.5}$	SUMMER	13.5	12.7	0.094	1.26	1.08	14.2	7.37
OB-7	$PM_{2.5}$	SUMMER	9.90	11.0	0.111	1.18	1.06*	10.3	7.05
OB-8	PM_1	SUMMER	6.20	9.26	0.149	0.95	0.93	1.37	21.2
OB-9	PM_1	SUMMER	7.10	7.49	0.105	0.90	0.92	2.12	17.9
OB-10	PM_1	SUMMER	9.50	8.70	0.092	1.01	1.09	8.34	13.8
OB-11	$PM_{2.5}$	FALL	34.3	35.7	0.104	2.99	2.54	15.0	33.7
OB-12	$PM_{2.5}$	FALL	14.6	35.9	0.246	1.47	1.25	14.8	32.7
OB-13	$PM_{2.5}$	FALL	27.0	44.0	0.163	1.27	1.17	7.70	35.2
OB-14	PM_1	FALL	14.7	24.3	0.165	1.19	1.13	4.93	30.7
OB-15	PM_1	FALL	8.30	16.9	0.203	0.97	0.88*	9.10	21.6
OB-16	PM_1	FALL	15.3	37.1	0.242	1.14	1.16*	1.62	36.5
OB-17	$PM_{2.5}$	WINTER	38.0	74.5	0.196	1.07	1.18	10.5	35.5
OB-18	$PM_{2.5}$	WINTER	22.0	34.7	0.158	0.96	0.86	10.2	27.1
OB-19	$PM_{2.5}$	WINTER	55.0	78.0	0.142	1.10	1.19	8.53	42.1
OB-20	PM_1	WINTER	28.0	43.9	0.157	1.04	1.18	13.6	34.0
OB-21	PM_1	WINTER	11.0	21.0	0.191	0.96	0.86	9.91	24.9
OB-22	PM_1	WINTER	29.0	37.0	0.128	1.14	1.16	2.15	33.4
ASC-1	$PM_{2.5}$	SUMMER	7.10	8.05	0.113	1.04	1.04	0.31	8.60
ASC-2	$PM_{2.5}$	SUMMER	4.60	6.77	0.147	0.96	0.94	1.90	6.79
ASC-3	$PM_{2.5}$	SUMMER	1.40	6.97	0.498	1.00	1.05	4.84	4.27
ASC-4	$PM_{2.5}$	SUMMER	1.50	6.19	0.413	0.97	0.91	6.31	4.39
ASC-5	$PM_{2.5}$	SUMMER	3.10	6.20	0.200	1.01	1.08	6.52	4.67
ASC-6	$PM_{2.5}$	SUMMER	6.00	5.51	0.092	1.43	1.84	28.4	3.66
ASC-7	$PM_{2.5}$	SUMMER	8.20	3.73	0.045	1.67	2.05	22.9	1.76
ASC-8	$PM_{2.5}$	SUMMER	9.20	4.68	0.051	1.37	1.07	21.8	2.19
ASC-9	$PM_{2.5}$	SUMMER	1.50	4.97	0.331	1.29	1.06	17.6	1.35
ASC-10	$PM_{2.5}$	WINTER	2.84	6.32	0.223	0.88	0.93*	6.12	31.5
ASC-11	$PM_{2.5}$	WINTER	2.18	5.76	0.264	0.80	0.91*	14.3	18.4
ASC-12	$PM_{2.5}$	WINTER	2.06	11.7	0.566	0.82	0.93*	12.3	19.4
ASC-13	PM_1	WINTER	1.55	5.04	0.325	0.80	0.90*	12.9	28.9
ASC-14	PM_1	WINTER	1.18	6.24	0.528	0.77	0.88	14.7	24.9
ASC-15	PM_1	WINTER	0.93	1.56	0.167	0.97	0.95*	1.62	5.67
ASC-16	PM_1	WINTER	1.62	2.51	0.155	0.94	0.90	3.96	5.25

Table **S1.** Properties of the investigated PM samples, categorized according to the sampling site (MI, urban; OB, rural; ASC, remote sites, 1st column), particle size ($PM_{2.5}$, PM_1, PM_{10}, 2nd column) andseasonality (3rd column). The reported parameters are: PM concentration (ng/m^3, 4th column), n-alkane concentration (ng/m^3, 5th column) and relative value to the PM amount (6th column), CPI values computed according to the traditional procedure (CPI_{trad}, 7th column) and the $EACVF$ method (CPI_{EACVF} ,8th column), $\varepsilon\%$ (9th column) to estimate the relative estimation error of CPI_{EAVF} vs. CPI_{trad} values and $series\%$ (10th column) which is the relative contribution of the homologous series terms to the total amount of alkanes.

CPI_{EACVF} values marked by star in the 8th column indicate CPI_{EACVF} values computed from $EACVF_{res}$ using the complete procedure.

Data treatment

Theoretical models have been developed to express $EACVF$ in terms of the parameters describing the chemical composition of the analyzed sample [22, 23].
The value of $EACVF$ at the origin ($\Delta t = 0$) is expressed by the following equation:

$$EACVF(0) = \frac{A_T^2\left(\sigma_h^2/a_h^2 + 1\right)}{m_{tot} d_{h/2} 2.129X} \qquad \text{(S1)}$$

where A_T is the total area of the chromatographic signal, $d_{1/2}$ the half-height width of the $EACVF$ peak describing the mean separation performance, σ . The value σ_h^2/a_h^2 is the peak height dispersion ratio derived from the mean, a_h , and the variance, σ_h^2 , of peak heights computed from the separated peaks observed in the chromatogram.

If the signal contains an ordered sequence of peaks (following eq.1, main text) — generated by the terms of a homologous series contained in the sample — the computed $EACVF$ shows deterministic peaks located at interdistance $\Delta t = b$ and multiple values $\Delta t = kb$, where the value $\Delta t = b$ is the retention increment between the terms of the homologous series, e.g., the CH_2 retention time increment. The height of these peaks of the $EACVF$ plot, i.e., the values of $EACVF(kb)$, can be related to the properties if the series: the number of terms of the homologous series, n_{max} , and the odd/even distribution of the series terms [17, 23, 25].

The $EACVF_{tot}(kb)$ values computed for odd k values can be expressed by the following equation:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma}{X} 2a_{o,h}^2 a_{e,h}^2 \left[\frac{\sigma_h^2}{a_h^2} + 1\right](n_{max} - k) \qquad k=1,3,\ldots\ n_{max} - 1 \quad \text{(S2)}$$

The $EACVF_{tot}(kb)$ values computed for even k values are given by the following equation:

$$EACVF_{tot}(2b) = \frac{\sqrt{\pi}\sigma}{X}(a_{o,h}^2 + a_{e,h}^2)\left[\frac{\sigma_h^2}{a_h^2} + 1\right](n_{max} - 2) \quad k=2,4,\ldots\ n_{max} - 2 \ \text{(S3)}$$

where, in both the equations, σ_h^2/a_h^2 is the peak maximum dispersion ratio derived from the mean, a_h , and variance, σ_h^2 , of all the peak maxima computed from the separated peaks in the chromatogram.
The paremeters $a_{o,h}$ and $a_{e,h}$ are the mean peak heights computed on the odd and even terms of the series, $a_{o,h}$ and $a_{e,h}$, respectively: from them the R ratio is computed to estimate the CPI_{EACVF} value.

Algorithms based on eqs. S1-S3 have been developed to use the $EACVF$ values experimentally computed from the GC signal (eq.2) to estimate the properties of the separated mixture: the number of components, m_{tot} , the number of terms in the homologous series, n_{max} , the odd/even distribution described by CPI_{EACVF}, the contribution of homologous series described by $series\%$, and the contribution of UCM component [17, 23, 25].

Number of components, m_{tot}. From the $EACVF(0)$ value it is possible to estimate the total number of components present in the mixture, m_{tot}, and, from it, to calculate $a_T = A_T / m_{tot}$, i.e. the mean total chromatographic area.

Number of terms in the homologous series. The $EACVF_{tot}(bk)$ values computed for even k values at subsequent $\Delta t = kb$ and $\Delta t = b(k+2)$ interdistanes are used to estimate the number of terms n_{max}. In fact, starting from eq. S3, the following equation can be derived [17]:

$$n_{max} = 2\frac{EACVF_{tot}(bk)}{EACVF_{tot}(b(k+2))} + k \tag{S4}$$

Contribution of the homologous series. The $EACVF(2b)$ value, that is the value related to the total peak area of the series terms (($a_o^2 + a_e^2$) in eq. S3), can be compared to the $EACVF(0)$ value for the total chromatographic peak area ($A_T / m_{tot} = a_T$, in eq. S1) in order to estimate the relative contribution the homolous series terms make to the overall signal, according to the following expression:

$$series\% = \frac{EACVF(2b)}{EACVF(0)} \approx \frac{(a_o^2 + a_e^2)}{a_T^2} \tag{S5}$$

When $EACVF$ is computed on the GC-MS signal acquired in SIM mode at $m/z =$ 57+71+85 values, it selectively contains information on the alkane components in the mixture. Therefore, in this case, the $series\%$ parameter can be defined as an estimation of the relative contribution of the homologous series to the total alkane content of the sample.

Odd/even prevalence. By dividing eq. S2 for $k=1$ by eq. S3 for $k=2$, and introducing the R ratio to substitute $a_{o,h}$ and $a_{e,h}$, the following expression can be obtained as a function of R:

$$\frac{EACVF(b)}{EACVF(2b)} = \frac{\frac{2}{R}(n_{max}-1)}{\left(1+\frac{1}{R^2}\right)(n_{max}-2)} = \frac{2R(n-1)}{(R^2+1)(n-2)} \tag{S6}$$

This is a quadratic equation, that can be solved to obtain the R value directly from the whole chromatogram on which the $EACVF$ values are computed at $\Delta t = b$ and $\Delta t = 2b$.

Deconvolution of the UCM band component. If the GC signal contains a large UCM component contribution, the $EACVF$ computed on it shows the shape of UCM hump which is superimposed on the deterministic, n-alkane sequence peaks.

Bibliography

[1] T.A. Berger. *Chromatographia*, 42:63–71, 1996.

[2] I.D. Wilson, U.A.Th. Brinkman. *Journal of Chromatography A*, 1000:325–356, 2003.

[3] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen. *Journal of Chromatography A*, 1062:113–123, 2004.

[4] M.C. Pietrogrande, M.G. Zampolli, F. Dondi. *Analytical Chemistry*, 78:2579–2592, 2006.

[5] F. Dondi, M.C. Pietrogrande, A. Felinger. *Chromatographia*, 45:435–440, 1997.

[6] A. Felinger, M.C. Pietrogrande. *Analytical Chemistry*, 73:618A–626A, 2001.

[7] J.M. Davis, J.C. Giddings. *Analytical Chemistry*, 55:418–424, 1983.

[8] A. Felinger, L. Pasti, F. Dondi. *Analytical Chemistry*, 62:1846–1853, 1990.

[9] A. Felinger. *Data Analysis and Signal Processing in Chromatography.* Elsevier, Amsterdam, 1998.

[10] A. Felinger, L. Pasti, P.L. Reschiglian, F. Dondi. *Analytical Chemistry*, 62:1854–1860, 1990.

[11] A. Felinger, E. Vigh, A. Gelencsèr. *Journal of Chromatography A*, 839:129–139, 1999.

[12] A. Felinger, L. Pasti, F. Dondi. *Analytical Chemistry*, 64:2164–2174, 1992.

[13] M.C. Pietrogrande, F. Dondi, A. Felinger. *Journal of High Resolution Chromatography*, 19:327–332, 1996.

[14] F. Dondi, A. Betti, L. Pasti, M.C. Pietrogrande, A. Felinger. *Analytical Chemistry*, 65:2209–2215, 1993.

[15] A. Felinger, L. Pasti, F. Dondi. *Analytical Chemistry*, 63:2627–2633, 1991.

[16] M.C. Pietrogrande, L. Pasti, F. Dondi, M.H. Bollain Rodriguez, M.A. Carro Diaz. *Journal of High Resolution Chromatography*, 17:839–850, 1994.

[17] M.C. Pietrogrande, D. Ghedini, G. Velada, F. Dondi. *Analyst*, 123:1199–1204, 1998.

[18] M.C. Pietrogrande, P. Coll, R. Sternberg, C. Szopa, R. Navarro-Gonzalez, C. Vidal-Majar, F. Dondi. *Journal of Chromatography*, 939:69–77, 2001.

[19] M.C. Pietrogrande, I. Tellini, A. Felinger, C. Szopa, R. Sternberg, C. Vidal-Majar, F. Dondi. *Journal of Separation Science*, 26:569–577, 2003.

[20] M.C. Pietrogrande, I. Tellini, A. Felinger, C. Szopa, R. Sternberg, C. Vidal-Majar, F. Dondi. *Journal of Chromatography A*, 1002:179–192, 2003.

[21] M.C. Pietrogrande, I. Tellini, C. Szopa, A. Felinger, P. Coll, R. Navarro-Gonzalez, R. Sternberg, C. Vidal-Majar, F. Raulin. *Planetary and Space Science*, 51:581–590, 2003.

[22] M.C. Pietrogrande, M.G. Zampolli, F. Dondi. *Ann. Chim. (Rome)*, 94:721–731, 2004.

[23] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, C. Szopa, R. Sternberg, A. Buch, F. Raulin. *Journal of Chromatography*, 1071:255–261, 2005.

[24] M.C. Pietrogrande, M.G. Zampolli, F. Dondi. *Annali di Chimica*, 94:2579–2592, 2004.

[25] J.C.J. Giddings. *Journal of Chromatography*, 703:3–15, 1995.

[26] N. Marchetti, A. Felinger, L. Pasti, M.C. Pietrogrande, F. Dondi. *Analytical Chemistry*, 76:3055–3068, 2004.

[27] C. Horvath B.L. Karger, L.R. Snyder. *An Introduction to Separation Science.* J. Wiley, New York, 1973.

[28] C. Giddings. *Unified Separation Science.* J. Wiley, New York, 1991.

[29] A. Skvortsov, B. Trathnigg. *Journal of Chromatography A*, 1015:31–42, 2003.

[30] J. Harangi. *Journal of Chromatography A*, 993:187–195, 2003.

[31] M.C. Pietrogrande, M. Mercuriali, L. Pasti. *Analytica Chimica Acta*, 594:128–138, 2007.

[32] M.C. Pietrogrande, M. Mercuriali, L. Pasti, F. Dondi. *Analyst*, 134:671–680, 2009.

[33] R. Ladji, N. Yassaa, A. Cecinato, B.Y. Meklati. *Atmospheric Research*, 86:249–260, 2007.

[34] J. Schwartz. *Environmental Research*, 64:36–52, 1994.

[35] D.W. Dockery, C.A. Pope. *Annu Rev Public Health*, 15:107–132, 1994.

[36] C.A. Pope and D.W. Dockery. *Air pollution and health.* San Diego, Academic Press, 1999.

[37] W. Welthagen, J. Schnelle-Kreis, R. Zimmermann. *Journal of Chromatography A*, 1019:233–249, 2003.

[38] D.M. Brown, M.R. Wilson, W. MacNee, V. Stone, K.Donaldson. *Toxicol. Appl. Pharm.*, 175:191–199, 2001.

[39] H.E. Wichmann, A. Peters. *Phil Trans R Soc Lond*, 358:2751–2769, 2000.

[40] A. Peters, H.E. Wichmann. *Am J Respir Crit Care Med*, 155:1376–1383, 1997.

[41] H.E. Wichmann, C. Spix, T. Tuch, J. Peel, G. Wölke, A. Peters, J. Heinrich, W.G. Kreyling, J. Heyder. *Health Effects Institute, Cambridge, MA*, Report 98, 2000.

[42] A.J. Ghio, R.B. Devlin. *Am. J. Respir. Crit. Care Med.*, 164:704–708, 2001.

[43] P.E. Tolbert, M. Klein, K.B. Metzger, J. Peel, W.D. Flanders, K. Todd, J.A. Mulholland, P.B. Ryan, H. Frumkin. *J. Expo. Anal. Environ. Epidemiol.*, 10:446–460, 2000.

[44] G.R. Cass. *Trends in Analytical Chemistry*, 17:356–366, 1998.

[45] B.R.T. Simoneit, A.I. Rushdi, M.R. Bin Abas, B.M. Didyk. *Environmental Science and Technology*, 37:16–21, 2003.

[46] J. Cyrys, M. Stölzel, J. Heinrich, W.G. Kreyling, N. Menzel, K. Wittmaack, T. Tuche, H.E. Wichmann. *The Science of the Total Environment*, 305:143–156, 2003.

[47] P. Saxena, L.M. Hildemann. *J. Atmos. Chem.*, 24:946–949, 1996.

[48] M.C. Jacobson, H.C. Hansson, K.J. Noone, R.J. Charlson. *Rev. Geophys.*, 38(2):267–294, 2000.

[49] M. Kanakidou, J.H. Seinfeld, S.N. Pandis, I. Barnes, F.J. Dentener, M.C. Facchini, R.V. Dingenen, B. Ervens, A. Nenes, C.J. Nielsen, E. Swietlicki, J.P. Putuad, Y. Balkanski, S. Fuzzi, J. Horth, G.K. Moortgat, R. WInterhalter, C.E.L. Myhre, K. Tsigaridis, E. Vignati, E.G. Stephanou, J. Wilson. *Atmos. Chem. Phys.*, 5:1053–1123, 2005.

[50] *NARSTO, Particulate Matter Science for Policy Makers: A NARSTO Assessment.* EPRI 1007735, 2003.

[51] J.H. Seinfeld, J.F. Pankow. *Ann. Rev. Phys. Chem.*, 54:121–140, 2003.

[52] A. Middlebrook, D.M. Murphy, S.-H. Lee, D.S. Thomson, K.A. Prather, R.J. Wenzel, D.-Y. Liu, D.J. Phares, K.P. Rhoads, A.S. Wexler, M.V. Johnston, J.L. Jimenez, J.T. Jayne, D.R. Worsnop, I. Yourshaw, J.H. Seinfeld, R.C. Flagan. *J. Geophys. Res.-Atmospheres*, 108:8424–8432, 2003.

[53] A. Middlebrook, D.M. Murphy, D.S. Thomson. *J. Geophys. Res.-Atmospheres*, 103:475–483, 1998.

[54] D.M. Murphy, D.S. Thomson, T.M.J. Mahoney. *Science*, 282:1664–1669, 1998.

[55] *IPCC, Climate Change 2001.* Cambridge University Press: New York, 2001.

[56] P. Saxena, L.M. Hildemann, P.H. McMurry, J.H. Seinfeld. *J. Geophys. Res.*, 100:18755–18770, 1995.

[57] R.J. Sheesley, J.J. Schauer, J.D. Hemming, S. Geis, M.A. Barman. *Environmental Science and Technology*, 39(4):999–1010, 2005.

[58] S.H. Chung, J.H. Seinfeld. *J. Geophys. Res.*, 107:4407–4412, 2002.

[59] J. Haywood, O. Boucher. *Rev. Geophys.*, 38(4):513–543, 2000.

[60] M.C. Facchini, M. Mircea, S. Fuzzi, R.J. Charlson. *Nature*, 401:257–259, 1999.

[61] Q. Zhang, D.R. Worsnop, M.R. Canagaratna, J.L. Jimenez. *Atmos. Chem. Phys.*, 5:3289–3311, 2005.

[62] F.M. Bowman, C. Pilinis, J.H. Seinfeld. *Atmospheric Environment*, 29:579–590, 1995.

[63] R.J. Barthelmie, S.C. Pryor. *Sci. Total Environ.*, 205:167–178, 1997.

[64] K. Matsumoto, H. Tanaka, I. Nagao, Y. Ishizaka. *Geophys. Res. Lett.*, 24:655–658, 1997.

[65] *IPCC (Intergovernmental Panel on Climate Change).* Cambridge University Press: New York, 1995.

[66] J.R. Brown, R.A. Field, M.E. Goldstone, J.N. Lester, R. Perry. *Sci. Total Environ.*, 177:73–84, 1996.

[67] A. Dyremark, R. Westerholm, E. Överik, J.-A. Gustavsson. *Atmospheric Environment*, 29:1553–1558, 1995.

[68] Y. Kawanaka, E. Matsumoto, K. Sakamoto, N. Wang, S.-J. Yun. *Atmospheric Environment*, 38:2125–2132, 2004.

[69] V.A. Isidorov. *Organic Chemistry of the Earth's Atmosphere.* Springer-Verlag: Berlim, 1990.

[70] S.N. Pandis, R.A. Harley, G.R. Cass, J.H. Seinfeld. *Atmospheric Environment*, 26A:2269–2282, 1992.

[71] J.F. Pankow. *Atmospheric Environment*, 28:185–188, 1994.

[72] J.F. Pankow. *Atmospheric Environment*, 28:189–193, 1994.

[73] J. Yu, R.C. Flagan, J.H. Seinfeld. *Environmental Science & Technology*, 32(16):2357–2370, 1998.

[74] A. Gogou, N. Stratigakis, M. Kanakidou, E.G. Stephanou. *Organic Geochemistry*, 25:79–96, 1996.

[75] W.F. Rogge, M.A. Mazurek, L.M. Hildemann, G.R. Cass, B.R.T. Simoneit. *Atmospheric Environment*, 27A:1309–1330, 1993.

[76] J.J. Schauer, W.F. Rogge, L.M. Hildemann, M.A. Mazurek, G.R. Cass, B.R.T. Simoneit. *Atmospheric Environment*, 30:3837–3855, 1996.

[77] J.J. Schauer, G.R. Cass. *Environmental Science and Technology*, 34:1821–1832, 1996.

[78] M.A. Mazurek, G.R. Cass, B.R.T. Simoneit. *Environmental Science and Technology*, 25:684–694, 1991.

[79] M.A. Mazurek, M. Mason-Jones, H. Mason-Jones, L.G. Salmon, G.R. Cass, K.A. Hallock, M. Leach. *J. Geophys. Res.*, 102:3779–3793, 1997.

[80] M.A. Mazurek. *Environmental Health Perspectives*, 110:995–1003, 2002.

[81] P.A. Cranwell. *Organic Geochemistry*, 3:79–89, 1981.

[82] E.D. Evans, E.E. Bray. *Geochimica and Cosmochimica Acta*, 22:2–15, 1961.

[83] G. Rieley, R.J. Collier, D.M. Jones, G. Eglinton. *Organic Geochemistry*, 17:901–912, 1991.

[84] J.I. Hedges, F.G. Prahl. *Organic Geochemistry: principles and applications*, volume 237–253. Plenum Press, New York, 1993.

[85] R.P. Eganhouse, I.R. Kaplan. *Environmental Science and Technology*, 16(3):180–186, 1982.

[86] K. Pendoley. *Marine Pollution Bulletin*, 24(4):210–215, 1992.

[87] W. Jeng. *Marine Chemistry*, 102:242–251, 2006.

[88] M. Brault, B.R.T. Simoneit, A. Saliot, J.C. Marty. *Organic Geochemistry*, 12:209–219, 1998.

[89] M.C. Kennicutt II, C. Barker, J.M. Brooks, D.A. DeFreitas, G.H. Zhu. *Organic Geochemistry*, 11:41–51, 1987.

[90] W.C. Qu, B. Xue, C.W. Su, S.M. Wang. *Hydrobiologia*, 581:89–95, 2007.

[91] M. Obermajer, K.G. Osadetz, M.G. Fowler, L.R. Snowdon. *Organic Geochemistry*, 31:959–976, 2000.

[92] C.M. Reddy, T.I. Eglinton, R. Palić, B.C. Benitez-Nleson, G. Stojanović, I. Palić, T.I. Djordjević, G. Eglinton. *Organic Geochemistry*, 31:331–336, 2000.

[93] M. Maffei, S. Badino, S. Bossi. *Journal of Biological Research*, 1:3–19, 2004.

[94] G. Eglinton, R.J. Hamilton, . *Leaf epicuticular waxes. Science*, 156:1322–1335, 1967.

[95] M.A. Sonibare, A.A. Jayeola, A. Egunyomi. *Biochemical Systematics and Ecology*, 33:79–86, 2005.

[96] J. Bendle, K. Kawamura, K. Yamazaki, T. Niwai. *Geochimica et Cosmochimica Acta*, 71:5934–5955, 2007.

[97] I.G. Kovouras, E.G. Stephanou. *Indoor Air*, 12:17–32, 2002.

[98] M.C. Pietrogrande, M. Mercuriali, M.G. Perrone, L. Ferrero, G. Sangiorgi, E. Bolzacchini. *Environmental Science and Technology*, Submitted:–, 2010.

[99] D. Schuetzle, D. Cronn, A.L. Crittenden, R.J. Charlson. *Environmental Science and Technology*, 9:838–845, 1975.

[100] D. Grosjean, K.V. Cauwenberghe, J.P. Schmid, P.E. Kelly, J.N. Pitts Jr. *Environmental Science and Technology*, 12:313–317, 1978.

[101] S.T. Hatakeyama, T. Tanonaka, J. Weng, H. Bandow, H. Takagi, H. Akimoto. *Environmental Science and Technology*, 19:935–942, 1985.

[102] K. Kawamura, K. Ikushima. *Environmental Science and Technology*, 27:2227–2235, 1993.

[103] J.D. Blando, B.J. Turpin. *Atmospheric Environment*, 34:1623–1632, 2000.

[104] P. Warneck. *Atmospheric Environment*, 37:2423–2427, 2003.

[105] K. Kawamura, O. Yasui. *Atmospheric Environment*, 39:1945–1960, 2005.

[106] J.M. Lightstone, T.B. Onasch, D. Imre. *Journal of Physical Chemistry A*, 104:9337–9346, 2000.

[107] K.F. Ho, S.C. Lee, J.J. Cao, K. Kawamura, T. Watanabe, Y. Cheng, J.C. Chow. *Atmospheric Environment*, 40:3030–3040, 2006.

[108] K. Kawamura, H. Kasukabe, L.A. Barrie. *Atmospheric Environment*, 30:1709–1722, 1996.

[109] V.-M Kerminen, C. Ojanen, T. Pakkanen, R. Hillamo, M. Aurela, J. Merilainen. *Journal of Aerosol Science*, 31:349–362, 2000.

[110] K. Kawamura, F. Sakaguchi. *Journal of Geophysical Research*, 104:3501–3509, 1999.

[111] R. SèmpèreK. Kawamura. *Atmospheric Environment*, 30:1609–1619, 1996.

[112] K. Kawamura, R. Sèmpère, Y. Imai. *Journal of Geophysical Research*, 101:18721–18728, 1996.

[113] J. Ray, S. McDow. *Atmospheric Environment*, 39:7906–7919, 2005.

[114] X. Bi, B.R.T. Simoneit, G. Sheng, S. Ma, J. Fu. *Atmospheric Research*, 88:256–265, 2008.

[115] C. Oliveira, C. Pio, C. Alves, M. Evtyugina, P. Santos, V. Goncalves, T. Nunes, A.J.D. Silvestre, F. Palmgren, P. Wahlin, S. Harrad. *Atmospheric Environment*, 41:5555–5570, 2007.

[116] H. Wang, K. Kawamura, K.F. Ho, S.C. Lee. *Environmental Science and Technology*, 40:6255–6260, 2006.

[117] M. Li, S.R. McDow, D.J. Tollerud, M.A. Mazurek. *Atmospheric Environment*, 40:2260–2273, 2006.

[118] Y.C. Li, J.Z. Yu. *Environmental Science and Technology*, 39:7616–7624, 2005.

[119] X.F. Huang, L.Y. He, M. Hu, Y.H. Zhang. *Environmental Science and Technology*, 43:4665–4671, 2006.

[120] I. Kourtchev, L. Copolovici, M. Claeys, W. Maenhaut. *Environmental Science and Technology*, 43:4665–4671, 2009.

[121] I. Kourtchev, J. Warnke, W. Maenhaut, T. Hoffmann, M. Claeys. *Chemosphere*, 73:1308–1314, 2008.

[122] K. Kawamura, Y. Imai, L.A. Barrie. *Atmospheric Environment*, 39:599–614, 2005.

[123] M. Mochida, A. Kawabata, K. Kawamura, H. Hatsushika, K. Yamazaki. *Journal of Geophysical Research*, 108:4193, 2003.

[124] Z. Yue, M.P. Fraser. *Atmospheric Environment*, 38:3253–3261, 2004.

[125] M. Lewandowski, M. Jaoui, T.E. Kleindienst, J.H. Offenberg, E.O. Edney. *Atmospheric Environment*, 41:4073–4083, 2007.

[126] M. Jaoui, T.E. Kleindienst, M. Lewandowski, E.O. Edney. *Analytical Chemistry*, 76:4765–4778, 2004.

[127] A. Plewka, D. Hofmann, K. Müller, H. Herrmann. *Chromatographia*, 57:S253–S259, 2003.

[128] C. Schummer, O. Delhomme, B.M.R. Appenzeller, R. Wennig, M. Millet. *Talanta*, 77:1473–1482, 2009.

[129] M.C. Pietrogrande, D. Bacco, M. Mercuriali. *Analytical and Bioanalytical Chemistry*, 396:877–885, 2010.

[130] K. Kawamura, I.R. Kaplan. *Analytical Chemistry*, 56:1616–1620, 1984.

[131] J.M. Halket. *Handbook of Derivatives for Chromatography.* Wiley, New York, 1993.

[132] M.C. Pietrogrande, M. Mercuriali, D. Bacco. *Air Pollution XVI, Skiathos - WITPress, Southhampton*, 2008.

[133] X. Bi, G. Sheng, P. Peng, Y. Chen, Z. Zhang, J. Fu. *Atmospheric Environment*, 37:289–298, 2003.

[134] M. Mandalakis, M. Tsapakis, A. Tsoga, E.G. Stephanou. *Atmospheric Environment*, 36:4023–4035, 2002.

[135] A. Cincinelli, M. Del Bubba, T. Martinelli, A. Gambaro, L. Lepri. *Chemosphere*, 68:472–478, 2007.

[136] B.R.T. Simoneit. *Atmospheric Environment*, 18:51–67, 1984.

[137] P. Kotianovà, H. Puxbaum, H. Bauer, A. Caseiro, I.L. Marr, G. Èík. *Atmospheric Environment*, 42:2993–3005, 2008.

[138] L.E. Yu, M.L. Shulman, R. Kopperud, L.M. Hildemann. *Environmental Science and Technology*, 39:707–715, 2005.

[139] J.J. Schauer, W.F. Rogge, L.M. Hildemann, M.A. Mazurek, G.R. Cass. *Atmospheric Environment*, 41:S241–S259, 2007.

[140] M.C. Pietrogrande, N. Marchetti, A. Tosi, F. Dondi, P.G. Righetti. *Electrophoresis*, 26:2739–2748, 2005.

[141] M.C. Pietrogrande, N. Marchetti, F. Dondi. *Journal of Chromatography B*, 833:51–62, 2006.

Acknowledgements

Per prima vorrei ringraziare mia mamma, per tutto. Per l'affetto, la disponibilità, la generosità, il sostegno, l'aiuto morale e fisico. Per essermi sempre stata vicino e per avermi sempre messo al primo posto. Per avermi capito e, soprattutto, per aver cercato di capirmi e aiutarmi anche quando io proprio non ne volevo sapere.
Ringrazio mio babbo per avermi sempre voluto bene e per avermi sostenuto anche senza essere sempre presente fisicamente, ma con il cuore. Ringrazio mia sorella Alice...che da 28 anni non è ancora riuscita a liberarsi di me.
Ringrazio mia nonna Angela e mia zia Margherita per aver sempre creduto nelle mie capacità e per avermi aiutato quando ne avevo bisogno.
Ringrazio la Prof.ssa Pietrogrande e il Prof. Dondi per avermi indirizzato, aiutato e guidato in questi anni.
Ringrazio il Prof. Zanghirati per il fondamentale aiuto informatico e tecnico.
Ringrazio il Prof. Schnelle-Kreis, il Prof. Zimmermann, la Dr.ssa Abbaszade e la Dr.ssa Goldberg per avermi fatto vivere una bellissima esperienza, sia a livello accademico che a livello umano.
Ringrazio la mia pluriennale compagna di studi e amica Dr.ssa Valentina Costa, per essere arrivata fino a qui con me, per essermi sempre stata di sostegno morale e per tutte le passeggiate di sfogo verbale (e gossip) fatte insieme.
Ringrazio la mia coinquilina d'ufficio Dr.ssa Catia Contado, per i consigli, i confronti e le risate continue...soprattutto sui suoi campi di fragole...(e scusa ancora per averti 'iniziata' a facebook!).
Ringrazio i miei colleghi dottorandi per esserci sempre sostenuti a vicenda: Dr. Dimitri Bacco, Dr.ssa Giulia Basaglia, Dr.ssa Marianna Nassi, Dr. Michele Orlandi, Dr.ssa Alessandra Vecchi.
Ringrazio in particolare i miei amici Luca, Enrico, TatoCuomo, Stefano SS, Jacopo e Marco per tutto quello che abbiamo passato insieme.
Ringrazio i miei 'nuovi' amici Norino, Claudio, Luca e Alessandro per avermi accolto con affetto (diciamo più o meno tutti, vero Onofrio?).
E ora...dulcis in fundo...ultimo ma non ultimo...ringrazio un piccolo orsetto partenopeo. Ringrazio Claudio Marseglia (alias **Mr.Clio**) per avermi fatto capire, nonostante la mia testa dura, cosa vuol dire immaginarsi un futuro con qualcuno, una vita insieme. Grazie Claudio per farmi sen-

tire tutto quello che sto sentendo ora. Pensavo non mi sarebbe capitata mai una cosa del genere e invece, all'improvviso, sei arrivato tu. Ancora non capisco come sia possibile che una persona fantastica come te sia al mio fianco...

Grazie per il tuo bellissimo sorriso. Grazie per tutti i nostri momenti insieme, per le serate e le nottate passate abbracciati, con te che mi guardi mentre dormo perchè "quando dormo sono stupendo e dolce, perchè non dico nulla...", con te che mi fissi i "padiglioni auricolari" pronto a morderli. Grazie per il modo in cui mi guardi e sorridi perchè in quel momento sei contento. Grazie per le faccie che fai quando ti faccio un regalo e grazie per avermi sempre dato mille idee regalo, molto velatamente, ovvio.

Grazie per le canzoni che mi hai cantato...tutte.

Grazie per le risate che mi fai fare con le tue battute e i sorrisi che mi fai spuntare per la tua tanto nascosta dolcezza.

Grazie per il tuo buonissimo pollo al curry, per le tue lasagne, per i tuoi biscotti, per tutto quello che hai sempre cucinato per me.

Grazie di tutto Claudio.....e soprattutto....grazie che mi SOPPORTI.

E infine, grazie a tutte le altre persone che mi hanno aiutato in questi anni e che non ho menzionato.