Talanta 83 (2011) 1225–1232

Contents lists available at ScienceDirect

Talanta

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

2D autocovariance function for comprehensive analysis of two-way GC–MS data matrix: Application to environmental samples

Maria Chiara Pietrogrande[a,*], Dimitri Bacco[a], Nicola Marchetti[a], Mattia Mercuriali[a], Gaetano Zanghirati[b]

[a] Department of Chemistry, University of Ferrara, Via L. Borsari, 46, 44100 Ferrara, Italy
[b] Department of Mathematics, Math. Tech. Center, University of Ferrara, Ferrara, Italy

ARTICLE INFO

Article history:
Available online 1 August 2010

Keywords:
GC–MS data matrix
2D autocovariance function
Signal processing
Environmental chemical tracers
Atmospheric aerosol

ABSTRACT

This paper describes a signal processing method for comprehensive analysis of the large data set generated by hyphenated GC–MS technique. It is based on the study of the 2D autocovariance function (2D-EACVF) computed on the raw GC–MS data matrix, extending the procedure previously developed for 1D to 2D signals. It appears specifically promising for GC–MS investigation, in particular to single out ordered patterns in complex data: such patterns can be simply identified by visual inspection from deterministic peaks in the 2D-EACVF plot.

A case of order along the retention time axis ($x = t_R$) is represented by a horizontal sequence of peaks, located at the same interdistance $\Delta t_R = b_x$, e.g., b_x is the CH_2 retention time increment between subsequent terms of an homologous series. The order along the fragment mass axis ($y = m/z$) contains information on analyte fragmentation patterns. Deterministic peaks appear in the 2D-EACVF plot at $\Delta m/z$ values corresponding to the most abundant ion fragments – dominating fragments in MS spectrum – or to ions generated by repetitive loss of the same ion fragment, i.e., $\Delta m/z = 14$ amu produced by the $[CH_2]^{\bullet}$ group loss in n-alkanes.

Method applicability was tested by processing GC–MS data of organic extracts of atmospheric aerosol samples: attention is focused on identifying and characterizing homologous series of organics, i.e., n-alkanes and n-alkanoic acids, since they are considered molecular tracers able to track the origin and fate of different organics in the environment.

1. Introduction

Gas chromatography coupled with mass spectrometry (GC–MS) is currently the most widely used technique for analyzing volatile organic pollutants in environmental samples. The very high number of applications is the result of the high efficiency of gas chromatography separation and the good qualitative information and high sensitivity provided by mass spectrometry [1–5]. When gas chromatography is combined with MS, it generates extensive amounts of data—2 or 3 orders of magnitude larger than those from conventional GC. However, chromatographic peak overlapping, always present when multicomponent samples are separated, makes interpretation of such data difficult as it affects the qualitative analysis by mass spectral information and worsens the quantitative measurement.

The quantity and complexity of GC–MS data make human analysis of signals difficult and time-consuming, thus making computer-assisted signal processing necessary to transform the data into usable information, in particular, to deconvolve incompletely resolved peaks and to interpret the chromatogram, extracting all the analytical information hidden therein [6–13].

This paper describes a method for analysis of second-order GC–MS data to provide a comprehensive picture of the data matrix as a whole. The method is based on the 2D autocovariance function (2D-ACVF) computed on the raw GC–MS data matrix: the procedure has been previously developed, and widely applied for mono-dimensional chromatograms [14–21] and has been further extended to 2D separations [22–25]. This extension to 2D GC–MS data matrix appears particularly promising since it allows identification of specific 2D data features hidden inside the data complexity, in particular ordered patterns can be singled out. This is due to the statistical basis of the approach: the total chromatogram is regarded as a statistical ensemble whose general attributes can be characterized, i.e., number of components, peak width, retention and abundance patterns, and extent of separation. This approach differs from deconvolution methods where a short section of the

[*] Corresponding author.
E-mail address: mpc@unife.it (M.C. Pietrogrande).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.07.056

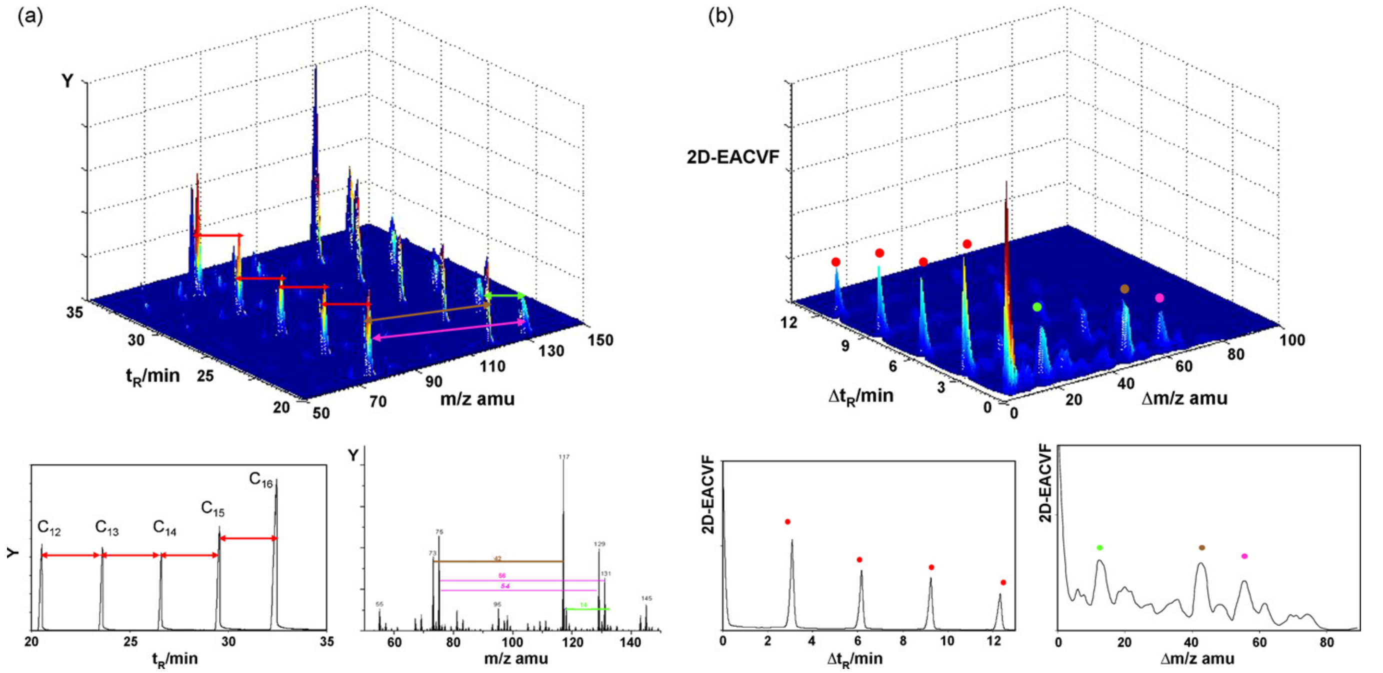

Fig. 1. GC–MS signal of a standard mixture containing C_{12}–C_{16} n-alkanoic acid silyl derivatives. (a) GC–MS data matrix of C_{12}–C_{16} n-alkanoic acids. Red arrows: constant interdistance $\Delta t_R = b_x$ between subsequent terms of the homologous series (SIM signal at $m/z = 75 + 117$ amu in the enlarged insert on the left). Green, brown and purple arrows: interdistance between the most abundant ion fragments at $m/z = 75$, 177 and 131 amu (MS spectra of C_{19} n-alkanoic acid in the enlarged inset on the right). (b) Plot of the 2D-EACV computed on the data matrix (a), positive quadrant: deterministic 2D-EACV peaks are identified by coloured points. Red points (corresponding to red arrows in (a)): constant interdistance values $\Delta t_R = b_x = 3.1$ min between subsequent terms of the homologous series (EACVF on the SIM signal at $\Delta m/z = 42$ amu in the enlarged inset on the left). Green, brown and purple points (corresponding to green, brown and purple arrows in (a)): deterministic peaks at $\Delta m/z = 14$, 42 and 56 due to interdistance between the most abundant ion fragmentations in mass spectra (EACVF of the MS spectra at $\Delta t_R = b_x$ in the enlarged inset on the right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

chromatogram is investigated, usually one cluster of several overlapping peaks, and an algorithm is used to estimate the profiles of the individual peaks [5,8–10,12,13].

In this paper, the method is applied to GC–MS data from organic extracts of atmospheric aerosol samples, to obtain specific information for identifying and characterizing homologous series of organics (i.e., n-alkanes and n-alkanoic acids), as relevant markers for source apportionment studies [26–30].

2. Theory

2.1. The 2D autocovariance function of GC–MS data matrix

When a raw data file is acquired in a GC–MS analytical run, the instrument returns two-way data, i.e., a matrix containing a mass spectrum for each scan. The matrix dimension is $N_x \times N_y$, where the columns represent the elution times t_R (N_x columns of mass spectra) and rows represent mass spectra for each acquisition point (N_y rows containing chromatograms for different m/z values). By way of example, a GC–MS data matrix is reported in Fig. 1a (GC–MS of standard mixture containing C_{12}–C_{16} n-alkanoic acid silyl derivatives).

On the experimental GC–MS data matrix acquired in digitized form (Fig. 1a), the 2D autocovariance function can be calculated according to the equation [22]:

$$\text{2D-EACVF}_{k,l} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x-k} \sum_{j=1}^{N_y-l} \left(f_{i,j} - \bar{f}\right) \left(f_{i+k,j+l} - \bar{f}\right) \tag{1a}$$

$$k = -N_x, \ldots, -1, 0, 1, \ldots, N_x \tag{1b}$$

$$l = -N_y, \ldots, -1, 0, 1, \ldots, N_y \tag{1c}$$

where $f_{i,j}$ represents the signal intensity at the point (i,j), while $\bar{f}$ is the average intensity calculated over all the points sampled.

N_x and N_y are the maximum spans of the t_R and m/z values over which 2D-EACVF is calculated. All the nodes of the matrix $N_x \times N_y$ are equally spaced. Each point used for computation (k and l interdistance values) can be converted into $\Delta x = \Delta t_R$ and $\Delta y = \Delta m/z$, on the basis of the sampling frequency in GC–MS data acquisition. The following relationships are used:

$$\Delta x = k\tau_x \tag{2a}$$

$$\Delta y = l\tau_y \tag{2b}$$

$$C\,(\Delta x, \Delta y)\,\tau_x\tau_y = \text{2D-EACVF}(k, l) \tag{2c}$$

where τ_x and τ_y are the inderdistances between subsequent points on the t_R and m/z axes, respectively. The 2D-EACVF computed over a data matrix (Fig. 1a) can be plotted *vs.* the Δt_R and $\Delta m/z$ values along the two coordinate axes, thus obtaining the 2D-EACVF plot: the positive quadrant for $\Delta x = \Delta t_R \geq 0$ and $\Delta y = \Delta m/z \geq 0$ is reported for the sake of simplicity (Fig. 1b), since 2D-EACVF exhibits a C_2 symmetry (Eq. (1a)), i.e., correlations in positions (Δx, Δy) and ($-\Delta x$, $-\Delta y$) are equal, which means that both positive and negative Δt_R and $\Delta m/z$ shifts give the same 2D-EACVF value [22]. The 2D-EACVF plot represents the correlations between positions of subsequent peaks along the retention axis and mass fragment m/z values in mass spectra.

The 2D-EACVF plot shows a main bidimensional Gaussian peak computed at the axis origin (i.e., $k = \Delta t_R = 0$ and $l = \Delta m/z = 0$): theoretical expressions have been developed for data matrix describing 2D separations in order to express 2D-EACVF in terms of the parameters for separations along the two axes, i.e., the number of single components (SC), m_{tot}, the SC peak standard deviation, σ, the function describing the SC retention pattern (interdistance model, IM) and the abundance distribution (abundance model, AM) [22–25].

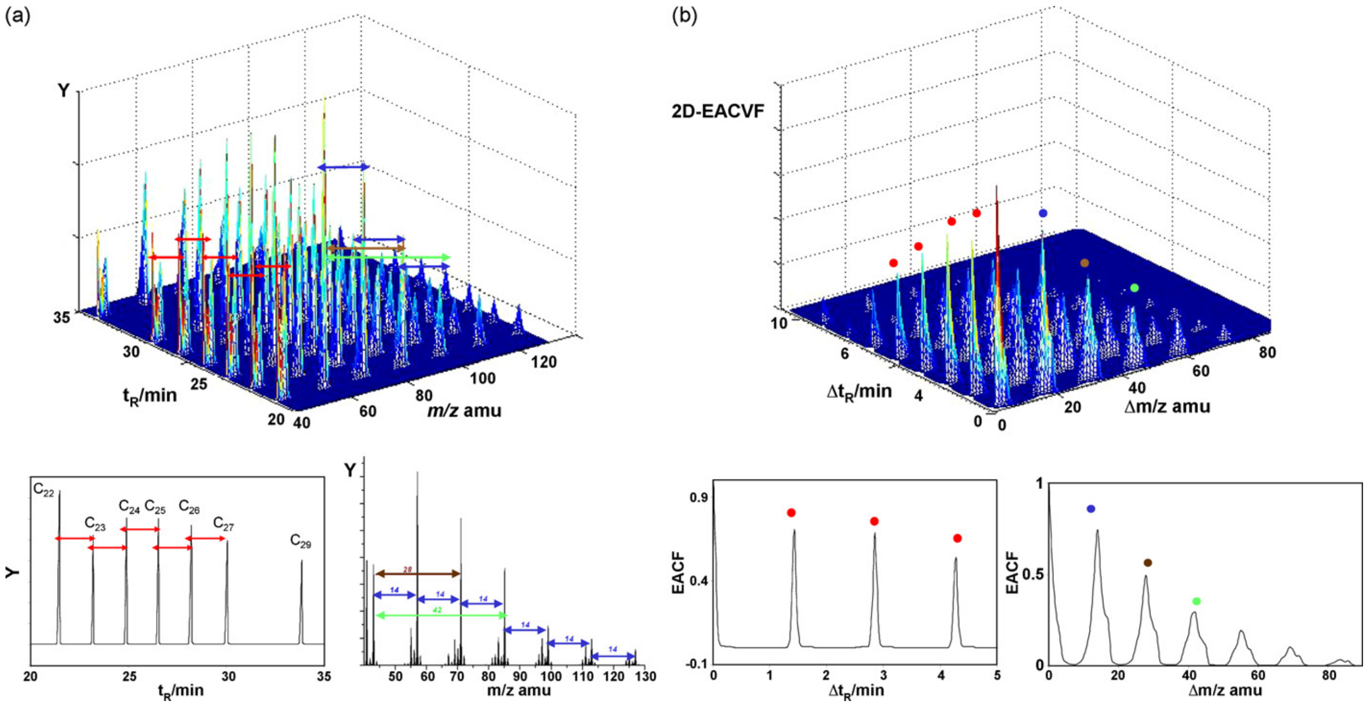

Fig. 2. GC–MS signal of a standard mixture containing C_{22}–C_{29} n-alkanes. (a) GC–MS data matrix of C_{22}–C_{29} n-alkanes. Red arrows: constant interdistance $\Delta t_R = b_x = 1.4$ min between subsequent terms of the homologous series (SIM signal at $m/z = 57 + 71 + 85$ amu in the enlarged insert on the left). Blue, brown and green arrows: interdistances between the most abundant ion fragments at $m/z = 14$, 28 and 42 amu (MS spectrum of C_9 n-alkane in the enlarged inset on the right). (b) Plot of the 2D-EACVF computed on the data matrix (a), positive quadrant: deterministic 2D-EACVF peaks are identified by coloured points. Red points (corresponding to red arrows in (a)): constant interdistance values $\Delta t_R = 1.4$ min between subsequent terms of the homologous series (EACVF on the SIM signal at $\Delta m/z = 14$ in the enlarged inset on the left). Blue, brown and green points (corresponding to blue, brown and green arrows in (a)): deterministic peaks at $\Delta m/z = 14$, 42 and 58 due to interdistance between the most abundant ion fragmentations in mass spectra (EACVF of the MS spectra at $\Delta t_R = 1.4$ min in the enlarged inset on the right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

In addition, the 2D-EACVF plot may show some peaks at $k = \Delta t_R$ and $l = \Delta m/z$ values (indicated by coloured points in Fig. 1b). These peaks at k,l correspond to the most abundant interdistances present in the original data, i.e., the most repeated ones or those between the highest peaks (coloured points in the 2D-EACVF plots in Figs. 1b and 2b). Therefore, they are called deterministic peaks, since they reflect the order in the data matrix: the presence of these ordered patterns can be simply identified by visual inspection of the 2D-EACVF plot, singling it out from the complexity of the 2D data [22–25].

In this paper, the 2D-EACVF method is applied for the first time to data matrix from a hyphenated technique where retention data, $x = t_R$, are coupled to MS data, $y = m/z$. The strength of the 2D-EACVF in identifying ordered sequences along the two-coordinate space is here applied to identify ordered structures independently generated along the two coordinate axes, i.e., retention time and mass spectra.

2.2. First dimension: retention times

GC signals from complex mixtures may display ordered structures or periodicities, due to the presence of sample components with related chemical structures; this order is difficult to be detected upon visual inspection, since it is usually superimposed to the random retention pattern that is the most common distribution in multicomponent samples [16–18]. An example of periodicity is described by the terms belonging to a homologous series that generates an ordered sequence of peaks with constant increments, while they do not necessarily display a substantial variation in m/z fragmentation pattern [18,19]. If $n_{\max}$ terms of the homologous series are present in the multicomponent sample, an ordered sequence of peaks will appear along the retention time axis ($x = t_R$), where the retention time of the nth term is described by:

$$x(n) = a_x + b_x n \qquad n = 0, 1, 2, 3 \ldots n_{\max} \tag{3}$$

where a_x represents the contribution of a specific functional group to overall retention, b_x the retention increment between subsequent terms of the homologous series, e.g., the CH_2 retention time increment ($\Delta t_R = b_x$ between subsequent terms of C_{12}–C_{16} n-alkanoic acids is indicated by red arrows in Fig. 1a). As another example, the GC–MS data matrix of a standard mixture containing C_{22}–C_{29} n-alkanes is reported in Fig. 2a: it shows a sequence of peaks, along the x axis, characterized by a constant interdistance $b_x = \Delta t_R = 1.4$ min between subsequent terms (indicated by red arrows).

2.3. Second dimension: MS spectra

The second dimension of GC–MS data matrix contains mass spectra, acquired in full scan mode. The fragmentation of molecular ions is a complex process including bond cleavage and molecular rearrangement to yield an assortment of fragment ions: their properties – m/z values and relative abundances – provide a clue to molecular structure and thus mass spectra are used as “fingerprints” to identify compounds [1–4]. Even if all sorts of fragmentations of the original molecular ion can produce ion fragments with random m/z value distribution, some bond cleavages yield more stable ion fragments and are to be preferred as they generate dominating fragments with higher abundance in the mass spectrum [3,5,10]. Moreover, organic compounds belonging to the same homologous series often exhibit similar fragmentation patterns with slight variations arising from the different substituents in the molecule. The result is that their mass spectra may contain characteristic ion fragments with specific m/z values diagnostic for the chemical class. An example is the MS spectrum of the trimethylsilylated derivatives of carboxylic acids showing dominating fragments at $m/z = 75$, 117 and 131 (MS spectrum of the sylilated C_{16} n-alkcanoic acid in the inset on the right of Fig. 1a) [15,31].

Moreover, some chemical structures may produce a fragmentation process characterized by repetitive loss of the same ion fragment. As an example, fragmentation pattern of n-alkanes is due to the C–C bond breaking to produce the repetitive loss of the $[CH_2]^\bullet$ group, corresponding to $\Delta m/z = 14$ amu. This is illustrated by the mass spectrum of C_9 n-alkane (enlarged detail on the right of Fig. 2a), where the propyl ($m/z = 43$ amu), butyl ($m/z = 57$ amu) and pentyl ($m/z = 71$ amu) ions are the most abundant fragments produced by the repetitive loss of $\Delta m/z = 14$ amu (blue arrows in Fig. 2a). In this case, the mass spectrum is formed by a sequence of signals located along the $y = m/z$ axis at the same interdistance $\Delta m/z = b_y$ so that the m/z value of the nth fragment is described by:

$$y(n) = a_y + b_y n \tag{4}$$

where b_x and b_y are constants.

2.4. 2D GC–MS data matrix

The 2D-EACVF computation achieves the comprehensive analysis of the GC–MS data matrix, since it simultaneously uses information on both retention and mass fragmentation, increasing the quantity and quality of information extracted from the data [5,10,12,13].

The combination of ordered distribution of signals independently located along the two coordinate axes, $x = t_R$ and $y = m/z$, generates an ordered pattern in the 2D matrix. In this case, the computed 2D-EACVF plot shows well defined deterministic peaks parallel to each coordinate axis located at the repeated Δt_R and $\Delta m/z$ values (coloured points in the 2D-EACVF plots in Figs. 1b and 2b) [22–25]. They can be simply identified by a visual inspection of the 2D-EACVF plot: the presence of these peaks is diagnostic for ordered structures and makes it possible to determine their b_x and b_y parameters (indicated by the coloured points in Fig. 2b corresponding to the coloured arrows in Fig. 2a). This property is the result of two concomitant abilities of the 2D-EACVF: it cancels the effect of position randomness while it amplifies the recursivity of the repeated interdistances [24].

To investigate specific pattern in detail, the 2D-EACVF can be projected on the coordinate axes (enlarged details of Figs. 1b and 2b). In particular, the intersection with the mass fragment axis ($\Delta m/z = 0$) describes the TIC chromatogram since it corresponds to the total MS signals acquired for each m/z value. 2D-EACVF intersections at specific $\Delta m/z$ values retain information on the SIM signals acquired at m/z values separated by the specific $\Delta m/z$ values: for n-alkanoic acids, 2D-EACVF intersection at $\Delta m/z = 42$ corresponds to the most abundant SIM signals at $m/z = 75$ and $m/z = 117$ (enlarged inset on the left of Fig. 1b). On these 1D chromatograms, the separation parameters can be quantitatively evaluated using the previously developed EACVF algorithms [16–19]. The appearance of deterministic EACVF peaks at $\Delta x = \Delta t_R = b_x$ and multiple values $\Delta t_R = b_x k$ is diagnostic for the presence of the series and their height (EACVF($b_x k$), i.e., the EACVF value computed at $\Delta t_R = b_x k$) is the basis for estimating the number of terms belonging to the ordered series, $n_{\max}$ [16,17]. Moreover, the EACVF values of subsequent peaks provide quantitative information on the distribution of the odd/even terms, quantified by the carbon preference index, CPI_{EACVF} [19].

The intersection of the 2D-EACVF plot with the retention time axis at Δt_R contains selective information on the molecular structure of the components. For example, for the n-alkanoic acids (GC–MS data in Fig. 1a) the 2D-EACVF section at $\Delta t_R = b_x = 3.1$ min (enlarged inset on the right of Fig. 1b) shows deterministic peaks at $\Delta m/z = 14$, 42 and 56 amu (coloured points in Fig. 1b and in the enlarged inset on the right). They are generated by the fragmentation pattern of the series terms, dominated by signals at $m/z = 75$, 117 and 131 amu values (coloured arrows in Fig. 1a and in the related enlarged inset on the right).

By combining qualitative information from the 2D-EACVF plot with quantitative data from the extracted 1D chromatograms, the GC–MS data can be chracterized in terms of retention behavior and mass fragmentation pattern.

3. Experimental

3.1. Chemicals

The standard mixtures containing known amounts of C_{19}–C_{33} n-hydrocarbons and n-alkanoic acids (from C_{12} to C_{26}) were prepared from standard compounds purchased from Fluka/Aldrich/Sigma (Sigma–Aldrich, Srl, Milan, Italy). The mixtures were prepared in methylene chloride by mixing proper amounts of standard compounds so that all terms of the homologous series display nearly the same concentration of 5 ng/μl.

The reagents used for derivatization of the carboxylic acids – BSTFA (bis(trimethylsilyl) trifluoroacetamide) plus 1% TMCS (trimethylchlorosilane) – were obtained from Aldrich Chemical Co. (Milan, Italy). All standards and reagents used were of the highest purity commercially available. All solvents were trace analysis grade (from 99.7%) from Sigma–Aldrich (Milan, Italy).

3.2. Instruments

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.

Proper temperature program conditions were selected for n-alkanes and n-alkanoic acids to obtain temperature programming conditions close to linearity, i.e., constant CH_2 retention time increments.

The temperature program for n-alkane analysis was set as follows: the initial temperature (120 °C) was held constant for 3 min and then raised to 295 °C at a rate of 5 °C/min, then further raised to 320 °C at 8 °C/min [19]. To analyze the carboxylic acids, the temperature-programmed analysis was performed by heating from 100 to 280 °C at a rate of 3 °C/min [32].

All samples were injected in split/splitless mode (mean split ratio: 1:20); the injector temperature was 300 °C.

The mass spectrometer operated in EI mode (positive ion, 70 eV): mass spectra were acquired in full scan mode with repetitive scanning from 40 to 400 m/z in 1 s. Ion source and transfer-line temperatures were 250 and 300 °C, respectively.

All the n-alkanes and n-alkanoic acids were identified by comparison with retention times and mass spectra of reference C_{19}–C_{33} n-alkane and C_{12}–C_{26} n-alkanoic acid standards.

3.3. Analytical procedure

The aerosol samples (PM_{10}) were collected daily on a pre-combusted quartz fiber filter (20 cm × 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 winter (January and February 2009) and in spring (April 2008) in different sampling sites close to Bologna (Northen Italy): urban (city centre of

Bologna) and rural sites (San Pietro Capofiume, a flat, homogeneous terrain of harvested fields, about 40 km north east of Bologna). After sampling, the procedure outlined in European Standard EN 12341 (CEN, 1998) was applied for equilibration and weighing.

The studied PM samples were extracted twice with 5 ml of dichloromethane (Sigma–Aldrich, Milan, Italy) using ultrasonic agitation for 20 min. The extracts were combined, filtered with a PTFE filter (0.45 μm) to remove insoluble particles. and then evaporated to dryness by a gentle stream of N_2. The sample was then dissolved in isooctane (50 μl) and directly injected into the GC–MS system for n-alkane determination [19].

For the n-alkanoic acid analysis, the dichloromethane extract was submitted to derivatization procedure: 30 μl of bis(trimethylsilyl) trifluoroacetamide (BSTFA) plus 1% trimethylchlorosilane (TMCS) were added to form trimethylsilyl (TMS) derivatives (reaction at 70 °C for 2 h). Then 2 μl of the sample was injected into the GC–MS system [32].

For the n-alkanoic acid analysis, the derivatization reaction was performed following the procedure reported in detail elsewhere [32]. The sample was transferred into a 1.5-ml tube and the solution evaporated to dryness. The silylation reagent was a mixture of BSTFA (bis(trimethylsilyl) trifluoroacetamide) plus 1% TMCS (trimethylchlorosilane): 30 μl of the reagent and 70 μl of n-isooctane were added into the tube, in addition to 5 μl of n-hexadecane used as an injection internal standard, IS (150 ng injected). The tube was sealed with a Teflon-coated cap and the reaction to form trimethylsilyl (TMS) derivatives was performed at the temperaure of 70 °C for the duration time of 120 min. Then 2 μl of the sample was injected into the GC–MS system.

3.4. Computation

The algorithms used for signal processing of GC–MS data and calculations were written in MATLAB© (The Mathworks, Inc. 2008) and run on a 1.53 GHz (256 RAM), AMD Athlon personal computer.

A new algorithm has been implemented to extend the original calculation on 1D chromatograms to a 2D data matrix [19]. The first step in GC–MS data handling was a linearization procedure to rescale the original t_R axis in order to obtain a strict constant retention increment $\Delta t_R = b_x$ between subsequent terms of the series. The terms of the homologous series were identified in the sample chromatogram, by comparison with retention times of standards analyzed under the same operating conditions; the Δt_R interdistance values between subsequent terms were evaluated, the maximum Δt_R was chosen as the reference value, $\Delta t_R = b_x$, and all the interdistances were stretched to reach the b_x value. An interpolation function was used in this stretching step to preserve the total area of the chromatogram.

The 2D-EACVF was then numerically calculated from the linearizated GC–MS data matrix according to Eq. (1a). A cyclic calculation procedure was used (the beginning and the end of the separation axes are merged using negative k or l indices): in this way, each point of the 2D-EACVF is computed using the same number of points and thus it is estimated with the same precision degree (see Ref. [22] for the details).

The previously developed MATLAB© algorithm was used to directly estimate the parameters m_{tot}, n_{max} and $\mathrm{CPI}_{\mathrm{EACVF}}$ from the 1D chromatograms obtained by intersection with the mass fragment axis ($\Delta m/z = 0$, TIC signal) and with specific $\Delta m/z$ values, i.e., SIM chromatograms [19].

4. Results and discussion

The applicability and usefulness of the 2D-EACVF method was tested on GC–MS data matrices obtained from the analysis of atmospheric aerosols (PM) and attention was focused on identification and characterization of homologous series present in the complex samples. This application is particularly relevant for environmental study, since it has been found that homologous series, i.e., n-alkanes and n-alkanoic acids, are especially suited to tracking the origin and fate of different samples, because they can originate from both man-made and natural sources and are highly resistant to biochemical degradation [26–30].

The applicability of the 2D-EACVF method was tested on PM samples with different chemical compositions stemming from variable contributions from different sources: three analyzed samples varied in seasonality – winter vs. spring (sample 1 vs. sample 2) – as well as sampling site location – urban vs. rural (sample 2 vs. sample 3).

4.1. n-alkanoic acid series

The 2D-EACVF method was applied to characterize n-alkanoic acids. It has been found that petroleum-based sources mainly emit low molecular weight n-alkanoic acids ($\leq C_{20}$), while plant waxes produce the heavier C_{20}–C_{30} terms. Moreover, the relative abundance of even-vs.-odd-numbered carbon n-alkanoic acids, described by the carbon preference index, CPI, is a key diagnostic parameter in tracking the biogenic vs. anthropogenic origin of organic inputs. In fact, anthropogenic emissions from fossil fuels generate a random distribution of even vs. odd terms yielding CPI values close to 1, while terrestrial plant material contains n-alkanoic acids with a predominance of even-numbered terms showing CPI ≥ 5 [27–30,33,34].

Analysis of these highly polar compounds requires a derivatization procedure prior to GC–MS analysis: a silylation procedure using BSTFA (bis(trimethylsilyl) trifluoroacetamide) and TMCS (trimethylchlorosilane) has been found suitable in analysis of the carboxylic acids in PM samples [15,32]. Moreover, the introduction of a silyl group (or groups) can enhance mass spectrometric signals of derivatives by producing a favourable fragmentation pattern dominated by derivatizing group fragments that are diagnostic for structure investigation. In fact, MS spectra of the TMCS derivatives are characterized by the most abundant fragments at $m/z = 73$ and 75, $[\mathrm{Si(CH_3)_3}]^+$ and $[\mathrm{HO{=}Si(CH_3)_2}]^+$, respectively, derived by substituting the active H atom with the $-\mathrm{Si(CH_3)_3}$ group. In addition, monocarboxylic acids show a strong fragment at $m/z = 117$, $[\mathrm{COOSi(CH_3)_3}]^+$, resulting from the trimethylsilyl group and acid functionality. Therefore, ions at $m/z = 73$ and 75 and 117 can be used to differentiate compounds bearing a –COOH group from other classes of organics (inset on the right of Fig. 1a: MS spectrum of the sylilated hexadecanoic acid) [31].

After derivatization, the PM samples were submitted to GC–MS analysis and 2D data matrix acquired: as an example the 3D plot of GC–MS data of sample 3 (collected in spring in a rural site) is reported in Fig. 3a (20–40 min region containing lighter C_{12}–C_{19} terms). An ordered sequence of peaks can be simply identified along the retention time axis (indicated by red arrows in the figure) and each peak displays the same fragmentation pattern along the m/z axis. The presence of the n-alkanoic acid homologous series can be confirmed by comparison with standards (GC–MS data of standard C_{12}–C_{16} terms under the experimental conditions, red arrows in Figs. 1b and 3b). All the studied samples show that the most abundant alkanoic acids are the C_{14}–C_{26} terms, with the highest contribution of hexadecanoic (C_{16}) and octadecanoic (C_{18}) acids [27,33].

This finding may be further supported by investigating the 2D-EACVF plot computed on the 2D data (Fig. 3b): it clearly shows deterministic peaks along the Δt_R and $\Delta m/z$ axes (Fig. 3b). Well-defined peaks are clearly evident at $\Delta t_R = 3.1$ min and multiple values, that are diagnostic for the presence of the homologous series

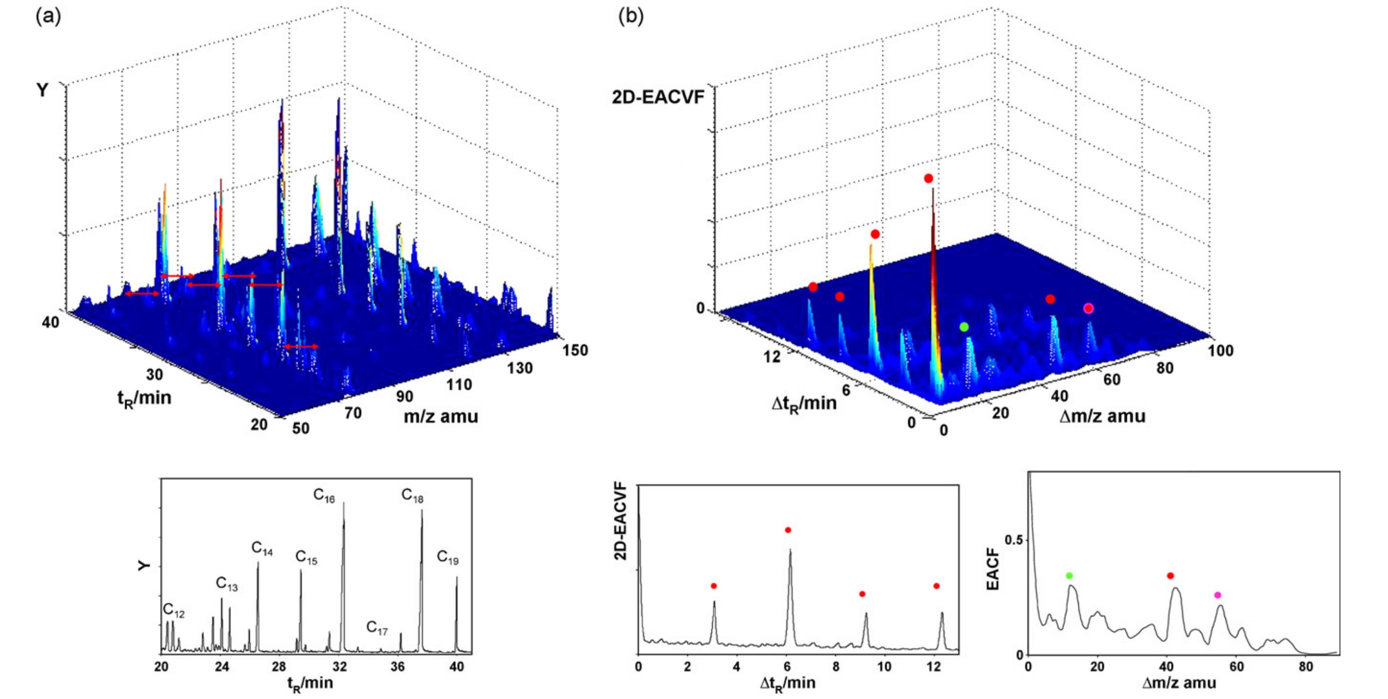

Fig. 3. GC–MS analysis of n-alkanoic acid silyl derivatives in PM sample 3. (a) GC–MS data matrix of the 20–40 min region containing lighter C_{12}–C_{19} terms. Coloured arrows: as in Fig. 1a (SIM signal at $m/z = 75 + 117$ amu in the enlarged insert on the left). (b) Plot of the 2D-EACVF computed on the data matrix (a), positive quadrant: deterministic 2D-EACVF peaks are identified by coloured points, as in Fig. 1b (EACVF on the SIM signal at $\Delta m/z = 42$ amu in the enlarged inset on the left). EACVF of the MS spectra at $t_R = 3.1$ min in the enlarged inset on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

(comparison with C_{12}–C_{16} standards, red points in Figs. 1b and 3b). In correspondence with these Δt_R values, deterministic peaks are present along the $\Delta m/z$ axis, that are diagnostic of n-alkanoic acid mass spectra: they are at $\Delta m/z = 14$, characteristic of the n-alkyl chain, and at $\Delta m/z = 42$ and 56, resulting from the differences in m/z values of the most abundant ions, i.e., $m/z = 75$, 117 and 131 (coloured points in Figs. 1b and 3b and in their enlarged insets on the right).

In addition, quantitative results can be computed by projecting the 2D-EACVF on the $\Delta m/z$ axis to extract 1D chromatograms, i.e., intersections at $\Delta m/z = 0$ to describe the TIC chromatogram and at $\Delta m/z = 42$ to represent SIM signals at the mass fragments characteristic of the series, i.e., $m/z = 75$ and $m/z = 117$ (enlarged inset on the left of Fig. 3b). From these 1D chromatograms, the separation parameters and the series properties can be quantitatively evaluated by the EACVF method: the total number of components, m_{tot}, from the TIC signal, the number of terms of the homologous series, n_{max}, and the carbon preference index, CPI_{EACVF}, from the signal at $\Delta m/z = 42$ [19]. Computations were performed on the three investigated samples and the results are reported in Table 1 (2nd–4th columns, EACVF estimation). The high CPI_{EACVF} values in the 6–9 range are consistent with the strong even/odd preference of n-alkanoic acids, in particular with the predominant contribution of the even terms C_{16} and C_{18}, that are known to be the most abundant species in most PM samples (enlarged detail on the left of Fig. 3a) [27–30,33]. In addition, the separation parameters were also estimated from the 1D chromatograms, TIC (intersection at $\Delta m/z = 0$, enlarged inset on the left of Fig. 3a) and SIM signal (inter-

Table 1
Properties of three investigated PM samples: comparison of data estimated using the 2D-EACVF method (2nd–4th columns, 2D-EACVF estimation) and conventional calculations on TIC and SIM chromatograms (5th–7th columns: conventional method).

n-alkanoic acids $CPI_{tot} = \Sigma(C_{14}–C_{26})/\Sigma(C_{13}–C_{25})$							
Sample	$\Sigma(C_{12}–C_{26})$ (ng/m^3)	2D-EACVF method			Conventional method		
		m_{tot}	n_{max}	CPI_{EACVF}	p_{tot}	n_{max}	CPI_{conv}
PM1 urban winter	134	26	13.1	6.5	20	13	6.3
PM2 urban spring	202	28	13.6	7.6	21	14	7.8
PM3 rural spring	225	29	14.2	9.8	21	14	9.2
n-alkanes $CPI_{tot} = \Sigma(C_{23}–C_{33})/\Sigma(C_{22}–C_{32})$							
Sample	$\Sigma(C_{21}–C_{33})$ (ng/m^3)	2D-EACVF method			Conventional method		
		m_{tot}	n_{max}	CPI_{EACVF}	p_{tot}	n_{max}	CPI_{conv}
PM1 urban winter	169.5	44	13.3	1.1	35	15	1.2
PM2 urban spring	17.6	42	13.1	1.2	32	13	1.2
PM3 rural spring	8.2	38	12.7	1.6	32	13	1.7

The reported parameters are: total concentration of homologous series terms (ng/m^3, 1st column), total number of components, m_{tot}, total number of n-alkanes, m_{hy}, and n-alkanoic acids, m_{ac}, number of homologous series terms present in the mixture, n_{max}, carbon preference index, CPI_{EACVF} and CPI_{conv} value, total number of peaks that can be counted in the TIC chromatograms, p_{tot}, and total number of alkanes, p_{hy}, and alkanoic acids, p_{ac}, that can be counted on the SIM chromatograms.

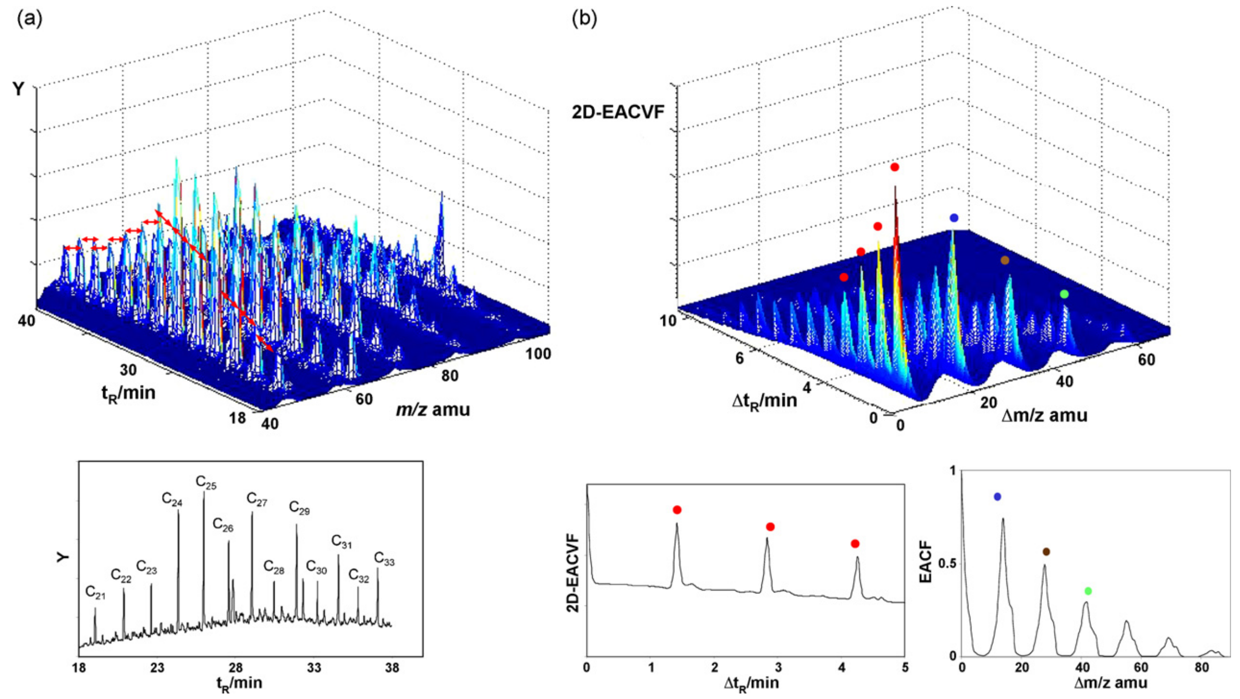

Fig. 4. GC–MS signal of the PM urban sample 1 for the analysis of C_{21}–C_{33} n-alkanes in PM sample 1. (a) GC–MS data matrix. Coloured arrows: as in Fig. 2a (SIM signal at $m/z = 57 + 71 + 85$ amu in the enlarged insert on the left). (b) Plot of the 2D-EACVF computed on the data matrix (a), positive quadrant: deterministic 2D-EACVF peaks are identified by coloured points, as in Fig. 2b (EACVF on the SIM signal at $\Delta m/z = 14$ in the enlarged inset on the left). EACVF of the MS spectra at $t_R = 1.4$ min in the enlarged inset on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

section at $\Delta m/z = 42$) using the conventional procedure, based on peak integration and computation on peak area. From the TIC chromatograms the number of peaks, p_{tot}, can be counted, while from the SIM signals the number of terms of the series, n_{max}, and their even/odd prevalence can be estimated, CPI_{conv} (obtained data in Table 1, 5th–7th columns, conventional calculations). The data obtained with the two independent procedures were compared to check the reliability of 2D-EACVF results. It must be underlined that the 2D-EACVF approach makes it possible to estimate the total number of components effectively present in the mixture – m_{tot} – while the parameter experimentally accessible with the conventional procedure is the number of peaks counted p_{tot}: usually $p_{tot} \leq m_{tot}$, as a consequence of peak overlapping.

The close agreement between the data (Table 1, n_{max}, 3rd vs. 6th columns, CPI_{EACVF}, 4th column, vs. CPI_{trad}, 7th column) is an experimental evidence that the study of 2D-EACVF and its intersections with the coordinate axes is indeed a useful tool for the comprehensive interpretation of the plethora of the GC–MS data, thus making it possible to obtain unequivocal identification of n-alkanoic acids and accurate characterization of their chemical composition.

4.2. n-alkane series

The PM samples were analyzed to extract information on the n-alkane content. The 3D plot of GC–MS data matrix obtained from the urban sample PM1 is shown in Fig. 4a, where the C_{21}–C_{33} n-alkane terms are identified. Their distribution profile shows the contribution of vehicular exhaust and lubricant residues, higher abundance of C_{24} and C_{25} terms, as well as emission from biological sources, characterized by C_{27}, C_{29}, and C_{31} terms, displaying odd carbon number preference (more clearly shown in the SIM chromatogram at $m/z = 57$, 71 and 85 amu, enlarged inset on the left of Fig. 4a). In addition, the GC–MS signal shows the presence of some alkanes that cannot be resolved by GC analysis and display a similar MS fragmentation pattern to generate a cluster of unresolved peaks (UCM band, in the enlarged inset of Fig. 4a). This is a typical pattern of GC analysis of PM samples, in particular those collected in urban sites, making the chromatogram evaluation challenging due to the convolution with the unresolved cluster [30,34–36].

The presence of a huge amount of n-alkanes in the sample generates an ordered pattern in the GC–MS signal formed by the sequence of n-alkane peaks located at a constant interdistance $\Delta t_R = 1.4$ min (red arrows in Fig. 4a and the enlarged inset on the left of Fig. 2a) along the t_R axis and an ordered series of mass fragments separated by a $\Delta m/z = 14$ amu increment along the mass axis (coloured arrows in the enlarged inset on the right of Fig. 2a).

As a consequence, the 2D-EACVF plot computed on the whole GC–MS data matrix (plot in Fig. 4b) shows well defined deterministic peaks located at $\Delta t_R = 1.4$ min and multiple values $\Delta t_R = 2.8$ min, $\Delta t_R = 5.2$ min along the Δt_R axis and at $\Delta m/z = 14$ amu and multiple values $\Delta m/z = 28$ amu, $\Delta m/z = 42$ amu along the $\Delta m/z$ axis (coloured points in Fig. 4b and its enlarged insets). These peaks are dignostic to identify the n-alkane series, even if they are superimposed to the contribution of the UCM band (comparison between Figs. 2b and 4b). This result confirms that the EACVF approach is a useful tool to deconvolve complex signals by separating the n-alkane ordered contribution from the unresolved cluster (compare enlarged insets on the left of Figs. 2b and 4b).

To quantitatively evaluate the separation parameters, the EACVF method was applied to 1D signals obtained by projecting the 2D-EACVF on the mass fragment axis at $\Delta m/z = 0$, i.e., TIC chromatograms, and at $\Delta m/z = 14$, i.e., SIM chromatograms, at $m/z = 57$, 71 and 85 amu. The separation parameters – the total number of components, m_{tot}, the number of terms of the homologous series, n_{max}, and the carbon preference index, CPI_{EACVF} – were computed and compared with the data obtained by the conventional procedure (Table 1, 2nd–4th columns, EACVF estimation, vs. 5th–7th columns, conventional calculations). The close agreement between the data (Table 1, n_{max}, 3rd vs. 6th columns, CPI_{EACVF}, 4th column, vs. CPI_{trad}, 7th column) is an experimental evidence of the reliability of the results obtained. In particular, CPI value can be accurately estimated, as a key parameter in differentiating between emissions stemming from anthropogenic use of fossil fuels – which generate a random distribution of odd vs. even terms yielding CPI values close to 1 – and those from terrestrial plant material containig a predominance of odd-numbered terms showing CPI ≈ 5–10 [27,30,35–39]. For all the studied samples, $\text{CPI}_{\text{EACVF}}$ values close to 1 were obtained suggesting a major contribution of petroleum residues derived from vehicular emissions as compared to biological inputs [35–38]. It must be noted that reliable results were also obtained from the GC–MS signals of urban PM samples (PM1 and PM2 samples), displaying the greatest interference from the unresolved UCM band [29,30,35–37]. Moreover, a close examination of the data in Table 1 points out that the method reliability is independent of the n-alkane content, since accurate results are obtained from samples largely varying in concentration level. In fact, the total n-alkane concentration in PM1 (urban winter sample) is nearly 20 times higher than in PM3 (rural spring sample, Table 1, 1st column) [29,34–37].

5. Concluding remarks

The described results reveal the effectiveness of the 2D-EACVF procedure for handling the tremendous amount of highly complex analytical-chemical information produced by hyphenated techniques. In particular, the strength of the 2D autocovariance function method lies in its ability to simply single out ordered structures hidden in the complex data, by combining information derived from retention behavior and mass fragmentation pattern. This has proved particularly useful in identifying and characterizing homologous series as molecular markers to trace the origin and fate of organics in the environment.

Therefore, the approach can be proposed as a signal processing method for direct data analysis of non-pretreated two-way data with a simple procedure, not very demanding in terms of computer power. This appears mainly promising for high-throughput analysis of the large data sets generated by chemical environmental monitoring, as a procedure able to extract large amounts of chemical information with low labour and time requirements.

As the first application to GC–MS data matrix, the described method only achieves a qualitative description of the 2D-EACVF plot, while the quantitative estimation of the separation parameters is based on the 1D chromatograms extracted from 2D data. Further developments of this topic are currently in progress and theoretical models and algorithms are being developed to extend the method to quantitative computations. The approach presented here can also be applied to other hyphenated separation techniques (e.g., HPLC–UV and LC–MS) where peak shift, peak shape changes and baseline contributions are often even bigger issues.

Acknowledgements

The authors thank F. Dondi and L. Pasti (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 personnel of ARPAER for their support in providing the aerosol samples.

References

[1] A. Cincinelli, M. Del Bubba, T. Martellini, A. Gambaro, L. Lepri, Chemosphere 68 (2007) 472–478.
[2] G. Wang, L. Huang, X. Zhao, H. Niu, Z. Dai, Atmos. Res. 81 (2006) 54–66.
[3] M.A. Mazurek, Environ. Health Perspect. 110 (2002) 995–1003.
[4] M. Mandalakis, M. Tsapakis, A. Tsoga, E.G. Stephanou, Atmos. Environ. 36 (2002) 4023–4035.
[5] L. Xu, L.-J. Tang, C.-B. Cai, H.-L. Wu, G.-L. Shen, R.-Q. Yu, J.-H. Jiang, Anal. Chim. Acta 613 (2008) 121–134.
[6] J.G. Shackmana, C.J. Watson, R.T. Kennedy, J. Chromatogr. A 1040 (2004) 273–282.
[7] J.H. Christensen, J. Mortensen, B.H. Asger, O. Andersen, J. Chromatogr. A 1062 (2005) 113–123.
[8] J. Krupcık, J. Mydlova, I. Spanik, B. Tienpont, P. Sandra, J. Chromatogr. A 1084 (2005) 80–89.
[9] J.M. Amigo, T. Skov, J. Coello, S. Maspoch, R. Bro, Trends Anal. Chem. 27 (2008) 714–725.
[10] C. Zeigler, K. MacNamara, Z. Wang, A. Robbat Jr., J. Chromatogr. A 1205 (2008) 109–116.
[11] G. Wang, Q. Ding, Z. Hou, Trends Anal. Chem. 27 (2008) 368–376.
[12] Z. Liu, W. Cai, X. Shao, J. Chromatogr. A 1190 (2008) 358–364.
[13] X. Shao, Z. Liu, W. Cai, Trends Anal. Chem. 28 (2009) 1312–1321.
[14] A. Felinger, M.C. Pietrogrande, Anal. Chem. 73 (2001) 618A–622A.
[15] M.C. Pietrogrande, I. Tellini, L. Pasti, F. Dondi, C. Szopa, R. Sternberg, C. Vidal-Madjar, J. Chromatogr. A 1002 (2003) 179–185.
[16] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, C. Szopa, R. Sternberg, A. Buch, F. Raulin, J. Chromatogr. A 1071 (2005) 255–261.
[17] M.C. Pietrogrande, M.G. Zampolli, F. Dondi, Anal. Chem. 78 (2006) 2576–2592.
[18] M.C. Pietrogrande, M. Mercuriali, L. Pasti, Anal. Chim. Acta 594 (2007) 128–138.
[19] M.C. Pietrogrande, M. Mercuriali, L. Pasti, F. Dondi, Analyst 134 (2009) 671–680.
[20] M.C. Pietrogrande, G. Basaglia, F. Dondi, J. Sep. Sci. 32 (2009) 1249–1261.
[21] M.C. Pietrogrande, G. Basaglia, J. Chromatogr. A 1217 (2010) 1126–1133.
[22] N. Marchetti, A. Felinger, L. Pasti, M.C. Pietrogrande, F. Dondi, Anal. Chem. 76 (2004) 3055–3068.
[23] M.C. Pietrogrande, N. Marchetti, A. Tosi, F. Dondi, P.G. Righetti, Electrophoresis 26 (2005) 2739–2748.
[24] M.C. Pietrogrande, N. Marchetti, F. Dondi, P.G. Righetti, J. Chromatogr. B 833 (2006) 51–62.
[25] F. Dondi, M.C. Pietrogrande, N. Marchetti, A. Felinger, in: S.A. Cohen, M.R. Schure (Eds.), Multidimensional Liquid Chromatography: Theory and Applications in Industrial Chemistry and the Life Sciences, John Wiley & Sons, Hoboken, NJ, USA, 2008, pp. 59–88.
[26] B.R.T. Simoneit, Int. J. Environ. Anal. Chem. 23 (1986) 207–237.
[27] J.J. Schauer, W.F. Rogge, L.M. Hildemann, M.A. Mazurek, G.R. Cass, B.R.T. Simoneit, Atmos. Environ. 30 (1996) 3837–3855.
[28] S.S. Park, M. Bae, J.J. Schaue, Y.J. Kim, S.Y. Cho, S.J. Kim, Atmos. Environ. 40 (2006) 4182–4198.
[29] M. Li, S.R. McDow, D.J. Tollerud, M.A. Mazurek, Atmos. Environ. 40 (2006) 2260–2273.
[30] X. Bi, B.R.T. Simoneit, G. Sheng, S. Ma, J. Fu, Atmos. Res. 88 (2008) 256–265.
[31] C. Schummer, O. Delhomme, B.M.R. Appenzeller, R. Wennig, M. Millet, Talanta 77 (2009) 1473–1482.
[32] M.C. Pietrogrande, D. Bacco, M. Mercuriali, Anal. Bioanal. Chem. 396 (2010) 877–885.
[33] C. Oliveira, C. Pio, C. Alves, M. Evtyugina, P. Santos, V. Goncalves, T. Nunes, J.D. Silvestre, F. Palmgren, P. Wahlinc, S. Harrad, Atmos. Environ. 41 (2007) 5555–5570.
[34] R. Ladji, R. Yassaa, C. Balducci, A. Cecinato, B.Y. Meklati, Atmos. Res. 92 (2009) 258–269.
[35] J.J. Lin, L.-C. Lee, Atmos. Environ. 38 (2004) 2983–2991.
[36] X. Bi, G. Sheng, P. Peng, Y. Chen, J. Fu, Atmos. Environ. 39 (2005) 477–487.
[37] Y. Cheng, S.-M. Li, A. Leithead, J.R. Brook, Atmos. Environ. 40 (2006) 2706–2720.
[38] A. Cincinelli, M. Del Bubba, T. Martinelli, A. Gambaro, L. Lepri, Chemosphere 68 (2007) 472–478.
[39] G. Caravaggio, J.-P. Charlang, P. MacDonald, L. Graham, Environ. Sci. Technol. 41 (2007) 3697–3701.