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, Ferrara University, Italy

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 the chemical signature of specific input sources – can be efficiently extracted from the complex GC-MS signal thus reducing the labour, 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

WIT Transactions on Ecology and the Environment, Vol 116, © 2008 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR080341

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, the key parameters of the homologous series – number of terms and the *CPI* value – 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*, that can be directly computed from the whole 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 Figure 1 shows the $EACVF_{tot}$ plot computed on the chromatogram of Figure 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 $EACVF_{tot}(0)$ value, 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 Figure 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 Figure 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*$_{tot}$*(bk)* 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:

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

where all the reported parameters can be directly estimated from the chromatographic signal: *X* is the total chromatogram time span, $\sigma^2{}_h/a^2{}_h$ is the peak height dispersion ratio describing the relative abundance distribution of the *n* terms of the series: it derives from the mean, $a^2{}_h$, and the variance, $\sigma^2{}_h$, 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:

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

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 series components, i.e., the C_{12}-C_{35} n-alkane range. Otherwise, the *CPI* index can be calculated on selected terms in order to describe specific source contribution to the sample, i.e., the *CPI*$_{plant}$ parameter is computed on the heavier C_{25}-C_{35} n-alkanes to describe the contribution of n-alkane 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 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 µL 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=0.25µm) (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 $m/z=57+71+85$ (Figures 1 and 2 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).

To extract information on the PM chemical composition, the $EACVF_{tot}$ was computed on the whole chromatographic signal ($EACVF_{tot}$ plots reported in insets of Figures 1 and 2: solid lines). The $EACVF_{tot}$ plots show well-defined deterministic peaks at Δt=1.9min and multiple values that are diagnostic for the presence of the n-alkane homologous series (b=1.9min in these experimental conditions). The number of n-alkanes present in the mixture, n, can be estimated from the $EACVF_{tot}$(1.9min) values (eqn (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} (eqn (2)): CPI_{tot}=1.1 and CPI_{tot}=1.6 were estimated for urban and rural samples,

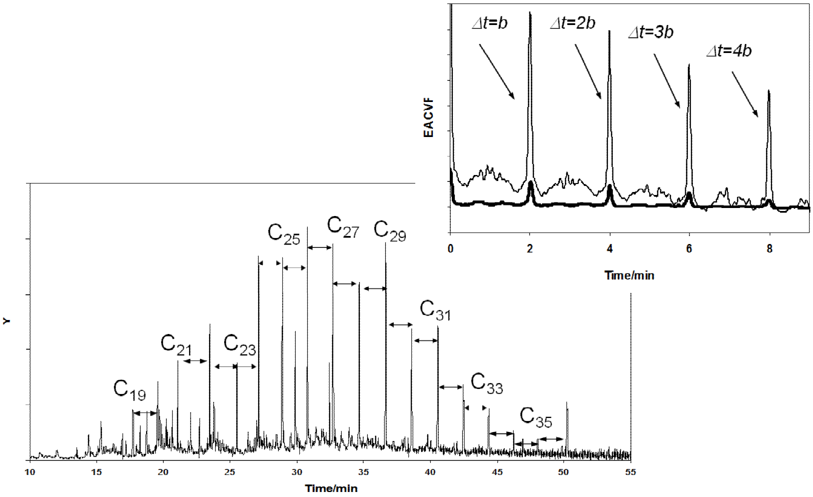

Figure 1: n-alkanes in urban $PM_{2.5}$: GC-MS chromatogram (SIM at *m/z=57+71+85*); inset: $EACVF_{tot}$ plot (solid line) and $EACVF_{plant}$ plot (bold line).

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 behaviour 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) (Figure 1): the $EACVF_{tot}$ of the urban sample (inset in Figure 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*=32-55min). For both samples, the number of terms n_{plant}=9 is estimated from the $EACVF_{plant}$ values at Δt=1.9min ($EACVF_{plant}$ plots in insets of Figures 1 and 2, 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 $EACVF_{plant}$ (inset of Figure 2, 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. The *EACVF* value at Δt=2*b*=3.8min is related to the total amount of the terms of homologous series (eqn. (1)): therefore, for each sample, the ratio between $EACVF_{tot}$(3.8min) and $EACVF_{plant}$(3.8min) 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.

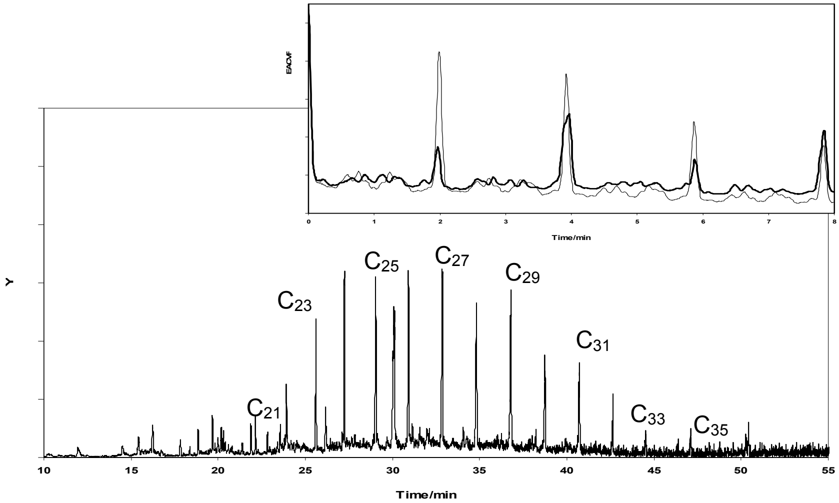

Figure 2: n-alkanes in rural $PM_{2.5}$: GC-MS chromatogram (SIM at *m/z= 57+71+85*); inset: $EACVF_{tot}$ plot (solid line) and $EACVF_{plant}$ plot (bold line).

To 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 marker 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

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

	EACVF Estimation				Traditional method			
Sample	*n*	CPI_{tot}	n_{plant}	CPI_{plant}	*n*	CPI_{tot}	n_{plant}	CPI_{plant}
n-alkanes CPI_{tot}=Σ(C_{13}-C_{35})/Σ(C_{12}-C_{34}); CPI_{plant}=Σ(C_{25}-C_{35})/Σ(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}=Σ(C_{14}-C_{30})/Σ(C_{13}-C_{29}); CPI_{plant}=Σ(C_{20}-C_{30})/Σ(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

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 by monitoring the typical fragments of the TMS derivatives at *m/z=75+147* (Figure 3a: rural sample). Under the experimental conditions used, the retention increment for subsequent n-alkanoic acids is *b*=2.5min. The $EACVF_{tot}$ was computed on the whole signal (Figure 3b: solid line): deterministic peaks at *Δt*=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 1, 1st-4th columns, *EACVF* estimation) and compared to results obtained with the traditional procedure (Table 1, 5th-8th columns, traditional calculations).

The $EACVF_{tot}$ plot shows a marked bimodal distribution with a predominant peak at *Δt*=2*b*=5min: 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 CPItot=6.9 values found for rural and urban samples, respectively (Table 1).

To extract information on the biological sources of n-alkanoic acids, the selected chromatographic region containing the C_{20}-C_{26} terms (35-60min) was separately investigated by computing $EACVF_{plant}$. The obtained $EACVF_{plant}$ plot

(Figure 3b, bold line) clearly identifies the contribution of biogenic sources, since it displays the strong bimodal distribution ($EACVF_{plant}(bk)$ peaks are low for $k=1, 3, 5$ and high for $k=2, 4$) 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 $EACVF_{tot}$(5min) and $EACVF_{plant}$(5min) 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.

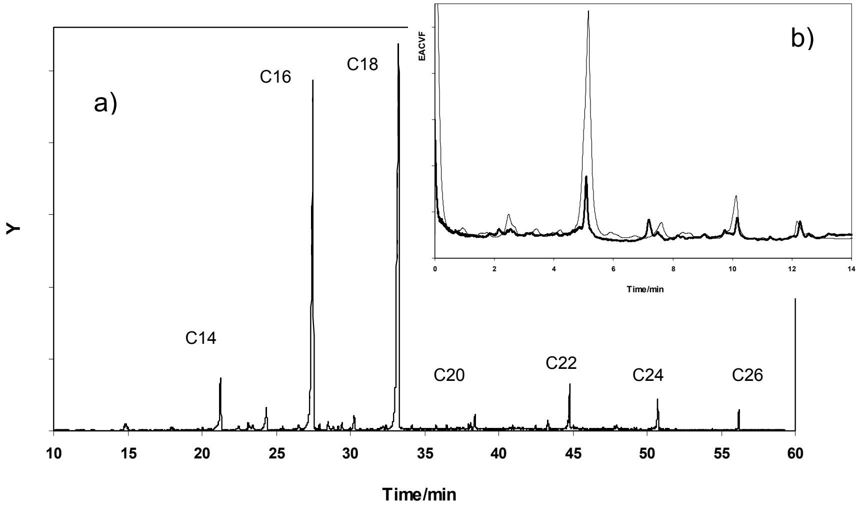

Figure 3: n-alkanoic acids in rural $PM_{2.5}$: a) GC-MS chromatogram (SIM at $m/z= 75+147$); b) $EACVF_{tot}$ plot (solid line) and $EACVF_{plant}$ plot (bold line).

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 labour 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.

Acknowledgement

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

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**, pp. 207–237, 1986.

[2] Schauer J.J., Rogge W.F., Hildemann L.M., Mazurek M.A., Cass G.R., Simoneit B.R.T., Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment,* **30**, pp. 3837–3855, 1996.

[3] Park S.S., Bae M., Schauer J.J., Kim Y.J., Cho S.Y., Kim S.J., Molecular composition of $PM_{2.5}$ organic aerosol measured at an urban site of Korea during the ACE-Asia campaign. *Atmospheric Environment,* **40**, pp. 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**, pp. 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**, pp. 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, *Atmospheric Environment,* **41**, pp. 5555–5570, 2007.

[7] Pietrogrande M.C., Zampolli M.G., Dondi F., Identification and Quantification of Homologous Series of Compound in Complex Mixtures: Autocovariance Study of GC/MS Chromatograms, *Anal. Chem.* **78**, pp. 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**, pp. 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.