Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1062 (2005) 113–123
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Chromatographic preprocessing of GC–MS data for analysis of complex chemical mixtures

Jan H. Christensen[a,b,*], John Mortensen[a], Asger B. Hansen[b], Ole Andersen[a]

[a] *Department of Life Sciences and Chemistry, Roskilde University, Universitetsvej 1, P.O. Box 260, 4000 Roskilde, Denmark*
[b] *Department of Environmental Chemistry and Microbiology, National Environmental Research Institute, Frederiksborgvej 399, P.O. Box 358, 4000 Roskilde, Denmark*

Received 27 May 2004; received in revised form 11 November 2004; accepted 12 November 2004

Abstract

Hyphenated analytical techniques such as gas chromatography–mass spectrometry (GC–MS) can provide extensive amounts of analytical data when applied to environmental samples. Quantitative analyses of complex contaminant mixtures by commercial preprocessing software are time-consuming, and baseline distortion and incomplete peak resolution increase the uncertainty and subjectivity of peak quantification. Here, we present a semi-automatic method developed specific for processing complex first-order chromatographic data (e.g. selected ion monitoring in GC–MS) prior to chemometric data analysis. Chromatograms are converted into semi-quantitative variables (e.g. diagnostic ratios (DRs)) that can be exported directly to appropriate softwares. The method is based on automatic peak matching, initial parameterization, alternating background noise reduction and peak estimation using mathematical functions (Gaussian and exponential-Gaussian hybrid) with few (i.e. three to four) parameters. It is capable of resolving convoluted peaks, and the exponential-Gaussian hybrid improves the description of asymmetric peaks (i.e. fronting and tailing). The optimal data preprocessing suggested in this article consists of estimation of Gaussian peak parameters and subsequent calculation of diagnostic ratios from peak heights. We tested the method on chromatographic data from 20 replicate oil samples and found it to be less time-consuming and subjective than commercial software, and with comparable data quality.

Keywords: Chromatographic peaks; Gaussian peak function; Exponential–Gaussian hybrid; Chemometrics; PCA; Chemical fingerprinting; Diagnostic ratios

1. Introduction

Hyphenated analytical techniques like gas chromatography–mass spectrometry (GC–MS), liquid chromatography–mass spectrometry (LC–MS), and gas chromatography–Fourier-transform infrared spectroscopy (GC–FTIR), are essential analytical tools for, e.g. research and development, environmental monitoring, and process chemistry. These methods have the capability of generating extensive amounts of data when applied to complex mixtures of contaminants as those present in polluted environmental samples (e.g. sediment, soil, sludge, and biota). Such samples often contain mixtures of contaminants with different physicochemical properties, degradability, and toxicity.

Despite of this inherent complexity, only a small percentage of the total number of compounds are usually considered for environmental monitoring and chemical fingerprinting [1–4]. A well-known example concerns monitoring and source correlation of polycyclic aromatic compounds (PACs), which are ubiquitous organic contaminants with varying toxicity, mutagenicity, and carcinogenicity. In addition to biogenic sources, PACs enter the environment from pyrogenic and petrogenic sources and their distributions are often very complex. However, most environmental studies include merely 16 PACs, selected from the US EPA priority pollutant list as relevant indicators of PAC pollution (http://www.epa.gov) [1–3,5]. Hence, in environmental monitoring and assessment studies, the number of target com-

* Corresponding author. Tel.: +45 46 30 12 00; fax: +45 46 30 11 14.
E-mail address: jch@dmu.dk (J.H. Christensen).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.11.037

pounds are often reduced prior to the chemical analysis. One reason is that data preprocessing is time-consuming and costly, but it also plays a role that univariate statistical analysis is vastly impeded when considering a large number of target compounds.

Chemometric methods, e.g. principal component analysis have been used frequently since the late 1990s for data analysis in environmental monitoring and chemical fingerprinting studies [3,5–7]. One advantage of multivariate compared to univariate statistical methods is the ease by which relations between multiple samples and variables can be resolved and visualized by score and loading plots. Additional advantages include noise reduction, obtained by multiple measurements of the same phenomenon (e.g. interrelated variables), and the ability to detect outliers [8]. However, multivariate methods still depend on chromatographic data preprocessing, which traditionally have focussed on resolving and quantifying peaks using internal and quantification standards.

In processing software included in software packages of mass spectrometers it is unlikely to select one set of peak identification and quantification parameters, optimal for hundreds of peaks with different signal-to-noise ratio, chromatographic resolution, and shape. Especially for baseline distorted and incompletely separated peaks, manual parameter adjustment is often necessary leading to increased subjectivity and time-consumption. The demands for quantification- and surrogate standards as well as retention time shifts are further impediments for preprocessing data from the analysis of complex chemical mixtures. Partly overlapping peaks are usually quantified using perpendicular-drop or tangent-skimming algorithms incorporated in commercial integrators [9,10]. However, these methods introduce systematic errors in the calculated peak areas and heights, depending on the degree of peak asymmetry and relative peak size [9–11]. Thus, to exploit the full potential of multivariate statistical methods for simultaneous statistical analysis of extensive data, and to reduce the subjectivity and uncertainty introduced by human intervention, there is a need to improve existing preprocessing methods.

One approach to deal with the above-mentioned problems is to perform the analysis on sections of digitized chromatograms [12–15]. However, such analyses are sensitive to even minute variations in retention times, because each individual scan number is a variable. Several alignment methods have been suggested to correct for retention time shifts in chromatographic data [13,15–17]. Dynamic time warping and correlation optimized warping are alignment methods that seem to work for a broad range of chromatograms [12,15,18–20]. Alignment combined with background noise reduction and normalization reduces the uncertainty and subjectivity introduced by peak quantification, and it makes simultaneous analysis of contaminant mixtures by, e.g. PCA feasible [12]. However, these methods can introduce erroneous data [19,20], and residual misalignment is sometimes present in data [12,19,20]. Furthermore, variations in peak shape (e.g. from symmetrical to tailing) during column deterioration will affect the multivariate data analysis negatively, due to changes in intensity distribution of adjacent retention times within a peak region. Peak quantification is less affected by these factors since peak areas and heights are relatively independent of peak shape.

Eide et al. (2001) presented a strategy to predict mutagenicity of organic extracts of exhaust particles from full scan GC–MS patterns of complex mixtures [21]. They based data preprocessing on an iterative curve resolution technique [22,23] capable of resolving second-order data (i.e. full-scan GC–MS data were resolved into chromatographic peaks and mass spectra). Curve resolution techniques have been used frequently to resolve overlapping peaks in second-order data [24,25]. However, these methods are not applicable to first-order data such as GC–MS with selected ion monitoring (SIM). Furthermore, iterative curve resolution techniques for resolving second-order data (GC–MS scan) are not likely to be able to resolve peaks in very complex chemical mixtures such as oil because multiple peak overlap (5–10 peaks) occur frequently.

Numerous mathematical functions have been used for presentation of chromatographic peaks and for deconvolution of incompletely resolved peaks in first-order data [26]. Many functions are based on the Gaussian function which gives good approximations of symmetric chromatographic peaks [27]. However, for asymmetric peaks, a Gaussian approximation is not adequate, and other more flexible functions give better peak descriptions [26–29]. On the other hand, these functions are often too flexible for proper approximation of experimental peaks, and are hence inapplicable for screening purposes.

In this paper, we present a method for semi-quantitative analysis and screening of complex chemical mixtures. The method is developed specific for semi-quantitative analysis of first-order data, and here we apply it for preprocessing GC–MS (SIM) data of petroleum hydrocarbons for use in forensic oil spill identification (i.e. source correlation). The method is based on automatic peak matching, initial parameterization, alternating background noise reduction and peak estimation, using mathematical functions with few parameters.

2. Method

The overall concept of a joint method for preprocessing first-order data from hyphenated analytical techniques is to reduce the time and cost of processing complex chemical data, increase data quality, and to increase objectivity. Our method consists of a collection of procedures that altogether convert chromatograms into semi-quantitative variables, i.e. diagnostic ratios (DRs), which have been used frequently for oil spill identification [30–33]. The method is capable of extracting ratios based on peak areas or heights, and the DRs can easily be externally normalized to the ratio in the laboratory reference analyzed closest in time to the analytical oil sample [31].

A compound database (i.e. retention times (RT), chemical names, and abbreviations) is set up prior to chromatographic preprocessing. Furthermore, semi-quantitative variables are defined in a variable database, which in contrast to the compound database can be redefined at all times during data processing. The preprocessing method is divided into a procedure for laboratory reference chromatograms, and one for sample chromatograms. The peak matching algorithm is based on the fact that chromatograms of replicate samples are identical, except for differences unrelated to the chemical composition (e.g. retention time shifts and detector sensitivity). Thus, peaks in replicate reference chromatograms are matched in a three-step procedure, whereas peaks in sample chromatograms are matched from retention times of the corresponding peaks in the reference analyzed closest in time. Subsequently, peak limits are determined as zero crossings of the first derivative of the chromatogram, on either side of the peak maximum. A subsection of these data points is used for calculating peak parameters, assuming either Gaussian or modified Gaussian peak shapes. The procedure combines background noise reduction (i.e. imposing an increasingly large noise limit to data) and simplex estimation of peak parameters. Finally, DRs and their uncertainties are calculated, quality controlled, and exported to chemometric software. Integer values (i.e. scan numbers) are used throughout the article as retention time. A flowchart of the method is shown in Fig. 1.

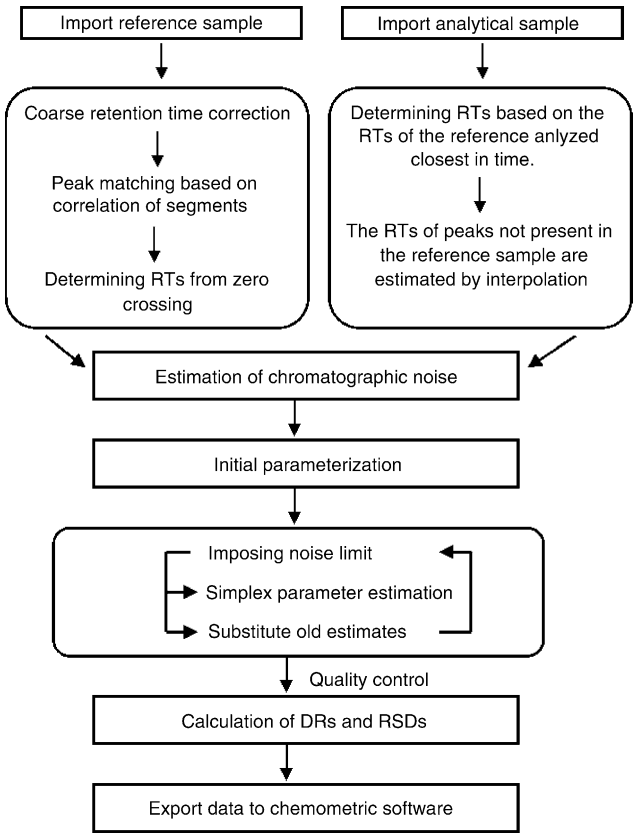

Fig. 1. Flowchart for the data preprocessing method.

2.1. Experimental and software

2.1.1. Application of a reference

Repeated analysis of a laboratory reference sample is a prerequisite for proper data preprocessing. Its sample characteristics and chemical composition need to be comparable to those of the analytical samples. Hence, the reference sample could be an authentic polluted environmental sample (for screening purposes) or a mixture of oil types (for oil spill identification). In the present work, a 1:1 mixture of Brent crude oil (North Sea crude) and a heavy bunker oil from the Baltic Carrier oil spill [31] was used as reference sample. Chromatographic data from replicate reference samples are used in the peak matching algorithm, but can also be used for external normalization of DRs, estimation of analytical uncertainty, and for quality control [31]. Reference samples need to be analyzed frequently depending on the rate of column deterioration. In this study, a reference sample was analyzed in the analytical sequence once per eight analytical samples.

2.1.2. Instrumentation and chemical analyses

Oil samples were analyzed on a Finnigan TRACE DSQ^{TM} Single Quadrupole GC–MS (Thermo Electron Corporation) operated in EI mode and equipped with a 60 m HP-5MS capillary column (0.25 mm i.d. × 0.25 μm film). One microliter aliquots were injected in PTV splitless mode. Starting temperature: 35 °C, increasing with 14.5 °C/s to 315 °C and hold for 1 min during transfer; carrier gas: helium (1.0 ml/min); temperature program: 35 °C (2 min), 60 °C/min to 100 °C, 5 °C/min to 315 (20 min); data acquisition time 2.34 scans/s. Mass spectrometric conditions: transfer line and ion source temperature: 300, and 250 °C, respectively. Forty-four mass fragments were analyzed in 8 groups of 12 ions using SIM. Twenty replicate reference samples analyzed over a period of 2 months were used to test peak matching, parameter estimates, and uncertainty of DRs.

2.1.3. Software

The data preprocessing method was implemented in Borland Delphi 4.0 object oriented programming. A version of the program can be obtained by contacting one of the authors. The network common data form (NetCDF) is an interface to a library of data access programs for storing and retrieving scientific data, developed by the Analytical Instrument Association (AIA). GC–MS chromatograms are exported to the AIA standard format, and NetCDF implemented in Matlab 6.5 is used to extract and sort data for import to the program. The NetCDF software is available for download at http://my.unidata.ucar.edu. The commercial GC–MS software Xcalibur 1.3 was used for comparative data preprocessing throughout the article.

2.2. Peak matching algorithm

The purpose of peak matching is to locate the position of a chromatographic peak (i.e. peak maximum) in reference and analytical sample chromatograms affected by run-to-run retention time variations. After the exact peak maximum has been found, the peak region (i.e. the data points part of the peak) can be determined, which is a prerequisite for the subsequent peak fitting.

2.2.1. Peak matching in reference chromatograms

The first part of peak matching in reference chromatograms is optional and consists of a manual time shift of each chromatogram by adding or subtracting a constant. Large constant retention time shifts can be generated by changing the chromatographic parameters (e.g. carrier gas flow, temperature program) or by cutting pieces of the capillary column.

2.2.2. Segment-wise peak matching

The fundamentals of the second part of the algorithm are that except for changes unrelated to chemical composition (e.g. retention time shifts, changes in sensitivity and peak shapes), replicate laboratory reference chromatograms should be perfectly correlated. Consider two chromatograms in which a number of peaks are to be matched. A peak (i) in one of these chromatograms is chosen as the target (T_i), and the peak maximum of the corresponding peak in the other reference chromatogram (P_i) is then matched to it, by comparing the appearance of the chromatogram surrounding the target peak (Fig. 2). Specifically, a segment of length M in the target chromatogram, centered at peak maximum (t_R), is compared to segments of equal length in the new reference chromatogram. For each peak matching, a finite number of possible integer shifts in the new reference defined by the maximum shift parameter (l) is investigated. The chromatographic features surrounding T_i is compared to those of 11 different segments when the maximum shift parameter is 5 (centered at −5 to +5 (Δ) of t_R).

The correlation coefficient ρ is used for chromatographic similarity matching of segments around T_i and P_i, respectively, because the degree of co-variation in segments indicates the quality of peak matching. The correlation coefficient (ρ) for two vectorized data sets (**a** and **b** of length M) is defined below (Eq. (1)) [34].

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\sum_{m=1}^{M}(a_m - \bar{a})(b_m - \bar{b})}{\sqrt{\sum_{m=1}^{M}(a_m - \bar{a})^2 \sum_{m=1}^{M}(b_m - \bar{b})^2}} \quad (1)$$

where M is the number of observations as well as segment length, and $\bar{a}$ and $\bar{b}$ the means of each data set. When using the correlation coefficient for chromatographic similarity matching, large peaks have a large influence on the variance compared to small ones, and thus the effect of noise is negligible. The optimal retention time shift (i.e. scan number) for P_i is the Δ-value with ρ closest to 1.

2.2.3. Determination of the peak maximum

In the third step of the algorithm, a more exact peak maximum is found. The algorithm initiates at the P_i found by segment-wise peak matching, and locates the zero crossing of the first derivative closest to P_i on either side. This value is the peak maximum in the new reference sample. A requirement for this step is that the residual shifts (between T_i and P_i), after segment-wise matching, are less than half the distance between neighboring peaks. Otherwise, peaks may be wrongly assigned. The estimated derivative is throughout the paper calculated for each point in a chromatogram as the difference between the intensity at the current scan number and the number that precedes it. In this way, integration is straightforward by cumulative summation.

2.2.4. Peak matching in analytical sample chromatograms

The retention time (i.e. scan number) of peaks in the laboratory reference sample analyzed closest in time to the considered analytical sample is used as initial estimates of t_R if peaks are present in the laboratory reference. If a peak is not present, t_R for that peak is estimated by polynomial interpolation along retention time using the shifts observed for other peaks in that chromatogram. In either case, the algorithm determines t_R as the zero crossing of the first derivative closest to this initial estimate. A requirement for correct peak matching in analytical samples is that for all peaks the

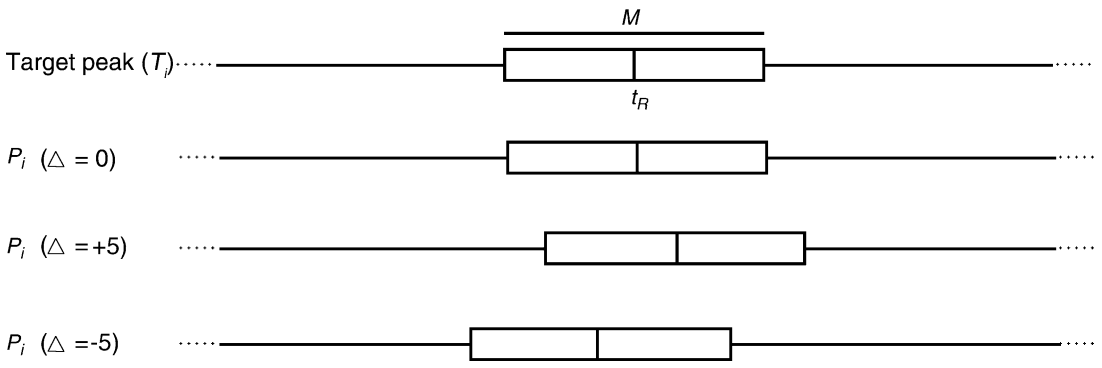

Fig. 2. Schematic presentation of segment-wise peak matching.

residual shift between the initial estimate of t_R and the peak maximum is less than half the distance between neighboring peaks.

2.3. Mathematical peak functions

The basic shape of chromatographic peaks is usually assumed to be symmetrical and can be approximated by a Gaussian distribution (Eq. (2)).

$$h(t) = H_R \exp\left(\frac{-(t - t_R)^2}{2\sigma^2}\right) \quad (2)$$

where t is the scan number, $h(t)$ the peak intensity, t_R, and H_R the position and height at the peak maximum, and σ the standard deviation of the Gaussian distribution. However, although a Gaussian distribution gives a reasonable fit to most experimental peaks, these are rarely symmetric. Skewing of peaks can occur by lengthening of either the leading or tailing edge, termed fronting and tailing, respectively [27]. Development of peak functions capable of describing asymmetrical peaks has been the subject of numerous publications [26–29,35], where e.g. the exponential modified Gaussian function (EMG) has received much attention [26,36].

The most frequent application of peak functions is for deconvolution of partially resolved peaks, and smoothing experimental peaks for the determination of statistical moments [26,37]. For these purposes, it is desirable that the mathematical models can describe peaks perfectly, and hence they should possess sufficient flexibility to fit peaks of different shapes. Consequently, peak functions found in literature are often flexible with five to nine parameters, but numerically unstable, and not easy to automatically fit to experimental peaks. For environmental screening purposes, a perfect peak fit is of less importance, and peak functions with acceptable fits to most experimentally occurring peak shapes, with the use of few parameters, are preferable. One such function, the exponential-Gaussian hybrid (EGH) was proposed by Lan and Jorgenson (2001) [38]. This function is closely related to the EMG, with higher stability and improved peak description at high asymmetries. Furthermore, the EGH defined in Eq. (3), has only one additional parameter (τ) compared to the Gaussian [38].

$$h(t) \equiv \begin{cases} H_R \exp\left(\dfrac{-(t-t_R)^2}{2\sigma_g^2+\tau(t-t_R)}\right), & 2\sigma_g^2 + \tau(t - t_R) > 0, \\ 0, & 2\sigma_g^2 + \tau(t - t_R) \leq 0, \end{cases} \quad (3)$$

where, σ_g is the standard deviation of the precursor Gaussian, and τ the time constant of the precursor exponential. Like the EMG, profiles of EGH approaches a Gaussian profile as $\tau \to 0$, and a truncated exponential profile as $\tau_g \to 0$. Notice that the EGH is easily replaceable by another function with few parameters.

2.4. Initial parameterization

An important feature of the Gaussian and the EGH model is that their parameters can be estimated from graphical information. We estimated initial values of t_R, H_R, and σ from the first and second derivatives of the chromatograms. t_R is the zero crossing of the first derivative using a first-order interpolation of the first positive and negative value on either side of the intersection. H_R is the peak height at this value, and σ equals half the peak width at the local extremes in the first derivative (i.e. zero crossings of the second derivative). The exact positions of the local extremes are determined by second-order fits using at least three scans surrounding the minimum and maximum, respectively. The mean of the two peak widths is used as an estimate of σ.

The time constant, τ of the precursor truncated exponential, can be estimated by [38]:

$$\tau = \frac{-(B_\alpha - A_\alpha)}{2\ln\alpha} \quad (4)$$

where A_α (tailing part) and B_α (leading part) are the distances from t_R, and α is the fraction of the peak height at which the distances are measured (e.g. $\alpha = 0.1$). We do not recommend that this procedure is applied to peaks unresolved at values above α, or baseline distorted peaks, since this may lead to erroneous estimations of A_α and B_α.

2.5. Simplex estimation of peak parameters

A simplex minimization procedure [39] was used for parameter estimation using either the Gaussian or EGH model, with the initial parameterization as starting values. t_R, H_R, σ, and τ were estimated for resolved peaks, whereas τ was fixed for incompletely resolved peaks. The skewness parameter (τ) for peak clusters may be approximated from resolved peaks in the neighborhood of the cluster, or from the fronting and tailing parts of the first and last peak in the cluster, respectively. The parameter estimations were terminated after a maximum number of iterations, or when the peak fit quality reached a lower limit.

2.6. Peak regions

Peak regions were determined from the peak maximum as zero crossings of the first derivative on either side of this value. Only data points within this region are used for initial parameterization and simplex estimation of peak parameters. For overlapping peaks, the peak region was defined as the start of the first peak to the end of the second peak. The Gaussian and EGH peak estimates utilize a different number of data points. The EGH utilizes data from the whole peak region, whereas the Gaussian utilizes data points within $\pm\sigma$ of the peak maximum.

2.7. Estimation of the chromatographic noise levels

In complex chemical mixtures, the chromatographic noise is composed of instrumental noise (electronic noise), and "chemical noise" (e.g. from co-elution of minor components). Especially, the latter varies between mass fragments, and along the retention time axis, and thus affects detection limits. The following iterative procedure was used to estimate chromatographic noise for each chromatogram.

(I) The standard deviation of the first derivative of a chromatogram is calculated.
(II) Peak regions containing at least one value larger than three times the standard deviation are left out and the standard deviation of the first derivative recalculated.
(III) Step II is repeated until only a certain percentage of data points, remains (default = 5%), no values exceed three times the standard deviation, or the number of iterations exceeds the maximum allowed (default = 35).

2.8. Peak estimation algorithm

An iterative peak estimation procedure based on chromatographic noise levels and the two mathematical peak functions is used to determine peak heights and areas. A lower (d_{min}) and upper (d_{max}) level of detection is defined as integer multiplications of the estimated noise for each chromatogram, and used to remove uninformative data (e.g. background and noise). The total number of steps (f), d_{min}, and d_{max} determines the step size (default values of d_{min} and d_{max} equal 1 and 20 times the noise, respectively, f equals 20, resulting in a step size of 1). The algorithm is elaborated below:

(I) The first derivative is calculated for each chromatogram.
(II) Setting the noise limit (increased for each step from d_{min} to d_{max} with step size).
(III) Data below current noise limit are set to zero, except for data points within peak regions, which are left unchanged.
(IV) Cumulative summation (i.e. integration) of the first derivative after imposing the current noise limit. Intensities are set to zero if these are less than zero after integration.
(V) Initial parameterization.
(VI) Simplex estimation of peak parameters and calculation of peak areas.
(VII) Estimated parameters substitute old ones and the algorithm continues from (II) if the peak area (or height) is smaller than or equal to the one calculated at lower noise limit, and at least one data point in the peak region is above the current noise limit. The iteration procedure is terminated if the peak area is equal to zero (i.e. peak below current noise limit).

The effects of increasing the noise limit are that more data points in the first derivative are set to zero, until at very high values where chromatographic information is no longer retained. Steps III and IV of the algorithm not only works by reducing the background outside peak regions, it also has an effect on the background noise within peak regions. Due to the decrease of the first derivatives outside peak regions by increasing the noise limit, the cumulative sums within peak regions are also reduced compared to the original chromatographic abundances, without subjective user interference. The purpose of the algorithm is thus to estimate peak parameters at some intermediate noise limit (the optimal). At this value, which varies for each peak, the background outside the peak region is set to zero, and data points within the region become background corrected.

3. Results and discussion

To obtain basic data for chromatographic preprocessing, a compound database was set up using the first reference sample in the analytical sequence. We selected 120 compounds from a suite of chemical groups of oil components listed in Table 1. The corresponding peaks represent a wide range of typical experimental peak shapes, from moderately fronting to tailing, well-separated, and incompletely resolved, as well as peaks with varying signal-to-noise ratios.

Table 1
Summary of the distribution of peaks within compound groups and corresponding m/z values

Compound group	m/z	Peaks	Compound group	m/z	Peaks
Naphthalene	128	1	Triterpanes	191	13
C_1-naphthalenes	142	2	Steranes/diasteranes	217/218	18
C_2-naphthalenes	156	8	Chrysene/benz(a)anthracene	228	2
C_3-naphthalenes	170	7	Triaromatic steroids	231	7
C_4-naphthalenes	184	9	Five ringed PAHs	252	3
Phenanthrene/anthracene	178	2	Fluorene	166	1
C_1-phenanthrenes/anthracenes	192	5	C_1-fluorenes	180	4
C_2-phenanthrenes	206	12	C_1-dibenzothiophenes	198	4
Dibenzothiophene	184	1	C_2-dibenzothiophenes	212	8
C_1-benzothiophenes	148	4	Deuterated surrogate mixture	136/212/264	3
C_1-pyrenes/fluoranthenes	216	6			

The variable database consisted of 74 DRs of single compounds. All ratios comprised peaks within the same chromatogram (i.e. *m*/*z* value) to decrease the variation caused by analytical uncertainty unrelated to the chemical composition [31]. The identities of DRs are not listed here, since they are irrelevant for testing the method. Twenty replicate reference samples analyzed as part of a large analytical sequence (i.e. 300 oil samples and quantification standards) were used for testing.

3.1. Peak matching

It was unnecessary to perform coarse retention time shifts for the current data set, since there were no large constant shifts between adjacent reference samples. The peak matching algorithm was used for matching peaks in 20 replicate reference chromatograms (P_i) by using the preceding reference in the analytical sequence as target (T_i) (e.g. the first was used as target for the second and so forth). Hence, the maximum shift parameter (l) depends on the maximum shift observed between adjacent reference samples in the analytical sequence. A peak consisted of approximately 15–25 data points depending on retention time and asymmetry. Segment sizes between 5 and 500 data points were tested, and for this data set, segments larger than 25 data points gave a perfect peak-match for all 120 peaks. For this study, we chose $l = 15$, which gave an adequate flexibility, and a segment size of 160 corresponding to the width of 6–10 peaks. Fig. 3 illustrates peak matching of an incompletely resolved peak, 2-/3-methyldibenzothiophenes. In Fig. 3b, segments of nine reference chromatograms have been shifted manually by adding or subtracting the optimal retention time shift for the specified peak. The quality of matching is evident, and similar results were observed for all 120 compounds in the peak database. Note that chromatograms are not retention time shifted as part of the complete preprocessing procedure. Conversely, the optimal retention time shifts are used for the subsequent peak fitting as an estimate of t_R.

The risk of wrong peak assignments increases for small segments and large maximum shifts. The former reduces the amount of chromatographic information available for similarity matching, whereas the latter increases the flexibility. On the other hand, it is a requirement that the flexibility is higher than the largest shift between adjacent samples. An upper limit of segment size for perfect peak matching was not reached for this data set, because run-to-run retention time variations were limited. For data containing large shifts and hence requiring a high flexibility, the use of large segments may result in bad assignments. The probability of incorrect peak assignments increases further when the signal-to-noise ratio for the specified peak is low, because its contribution to the correlation coefficient is small.

3.2. Iterative background noise reduction

Baseline subtraction is an important part of an automatic data preprocessing procedure. In standard chromatographic software, a baseline is set by the integration parameters, and it is often necessary to change these manually, especially for baseline distorted and incompletely resolved peaks. One way of automatically handling the baseline would be to perform polynomial- or piecewise-linear baseline fits. However, because peaks are often not baseline separated in complex chemical mixtures, it is difficult and subjective to determine baseline points.

In addition to iterative background noise reduction in the neighborhood of each selected peak, the procedure also defines detection limits. Fig. 4 illustrates the effects of iterative background noise reduction on six chromatographic peaks (tetracyclic steranes, *m*/*z* 217). The composition of steranes is often very complex with a highly elevated baseline and coeluting peaks. Hence, only a fraction of these com-

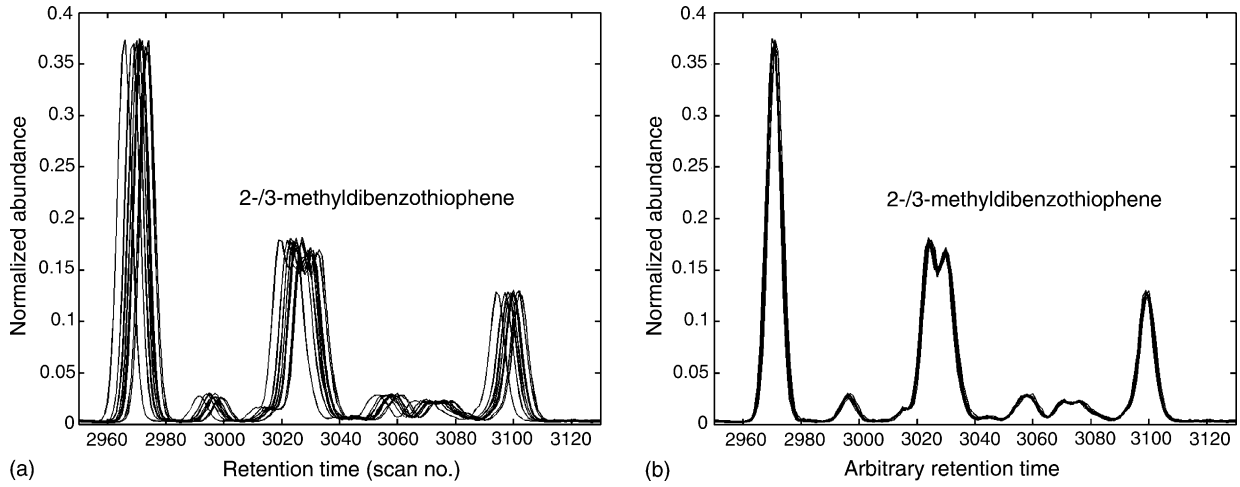

Fig. 3. Peak matching of chromatographic peaks. Section of the *m*/*z* 198 chromatogram for ten selected replicate laboratory reference samples: (a) before peak matching of 2-/3-methyldibenzothiophene (the two unresolved peaks); and (b) after manually shifting the chromatograms by adding or subtracting the optimal retention time shift for the specified peak. A segment length of 160 data points and a maximum shift of 15 were used in the peak matching algorithm. The data in the figure have been normalized to the average intensity to ease comparisons.

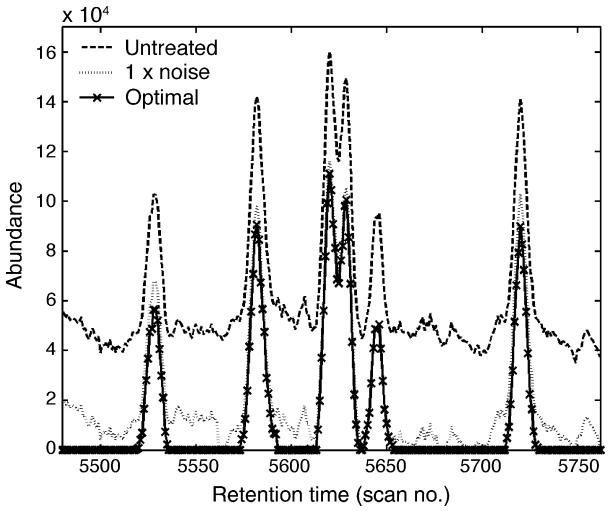

Fig. 4. Background reduction. A section of the *m/z* 217 chromatogram (tetracyclic steranes) from scan number 5480 to 5762 is shown for: I, the untreated chromatogram; II, after imposing a noise limit of 1 × noise; and III, after iterative noise removal with d_{min} = 1 × noise and d_{max} = 20 × noise using 20 steps.

pounds is commonly employed for chemical fingerprinting [32,40].

A significant part of the background is removed by imposing a noise limit of one times the noise. However, it is evident that the background noise reduction becomes more efficient when the iterative peak estimation algorithm is applied. This procedure employs the optimal noise limit for each of the six individual peak regions. The background correction seen here is comparable to that of commercial softwares, for peaks with large signal-to-noise ratios. Conversely, it is more objective and less time consuming for baseline distorted peaks with low signal-to-noise ratio.

3.3. Peak fit quality of single peaks

The simplex estimation procedure attempts to improve an initial set of peak-shape parameters by direct minimization in error space. Hence, its success in generating good parameter estimates is strongly dependent on the model function, how well it represents real peaks, and the method for initial parameterization. The relative root mean squared error (RRMSE) is used to evaluate the quality of each mathematical function for describing chromatographic peaks (Eq. (5)).

$$\mathrm{RRMSE} = 100 \times \sqrt{\frac{\sum_{n=1}^{N}(a_n - a_n^*)^2}{\sum_{n=1}^{N}(a_n)^2}} \quad (5)$$

where N is the number of data points in the peak region, a_n experimental observations, and a_n^* model estimations. Data points used for calculating RRMSE are shown as filled circles in Figs. 5 and 6. Fig. 5 illustrates how well the Gaussian and the EGH functions describe experimental peaks with a broad range of asymmetries, from moderately fronting to tailing (τ = −0.97 to +0.75).

The EGH gives a better description of the asymmetric peaks with RRMSE between 1.6 and 5.4% compared to 4.7 and 11.7% for the Gaussian function. Conversely, the two functions describe symmetric peaks, such as phenanthrene (Fig. 5c), equally well (RRMSE = 1.1% in both cases). For screening purposes, the fit quality is of less importance compared to systematic discrepancies in peak areas or heights. The 120 peaks considered in this study cover a broad range of asymmetries and signal-to-noise ratios. Data calculated by the Gaussian function deviated from those calculated by the EGH by between −9 and +2% for peak areas, and −2 to +4% for heights. These variations are systematic and depend on peak asymmetry related to the compound properties (e.g. boiling point and polarity). However, DRs are affected in the same direction when calculated for different reference and analytical samples, which minimize the significance of these systematic discrepancies.

3.4. Peak fit quality of incompletely resolved peaks

Commercial softwares use the perpendicular-drop-down or tangent-fitting methods to analyze incompletely separated peaks in commercial software. Since the EGH function contains an additional parameter compared to the Gaussian, it is more flexible in a way which affects the approximation of multiple peaks negatively (e.g. for two peaks eight parameters need to be approximated). Hence, for computational reasons and to avoid local minima, τ is fixed prior to the iteration, or a purely Gaussian description is applied. Here, we chose the Gaussian description (τ = 0) since the estimation of τ is difficult and uncertain when peaks are baseline distorted. Fig. 6 shows two examples of incompletely resolved peaks with height ratios of approximately 1:1 and 1:8, respectively.

The data set comprises a total of 10 peak clusters of incompletely resolved peaks with chromatographic resolution (R) < 1.5. $R = \Delta t/4\sigma$, where Δt is the difference in the retention time maxima of two components, and σ the average standard deviation of two Gaussian peaks. The iterative estimation of Gaussian parameters gives acceptable peak fits for peak clusters analyzed in this study, which include height ratios between 1:1 and 1:8, and chromatographic resolution as low as 0.65 (2-/3-methyldibenzothiophene in Fig. 4).

3.5. Quality of diagnostic ratios

The variability of DRs within 20 replicate reference samples was compared for Gaussian, EGH, and peak quantification using commercial software. DRs were calculated by Eq. (6).

$$\mathrm{DR}^{\mathrm{R}} = \frac{a_n^{\mathrm{R}}}{(a_n^{\mathrm{R}} + a_{n^*}^{\mathrm{R}})} \quad (6)$$

where a_n^{R} is the area or height of peak n or n^* in the sample. DRs were calculated for individual peak areas or heights (i.e. the area or height of the first peak divided by that of the

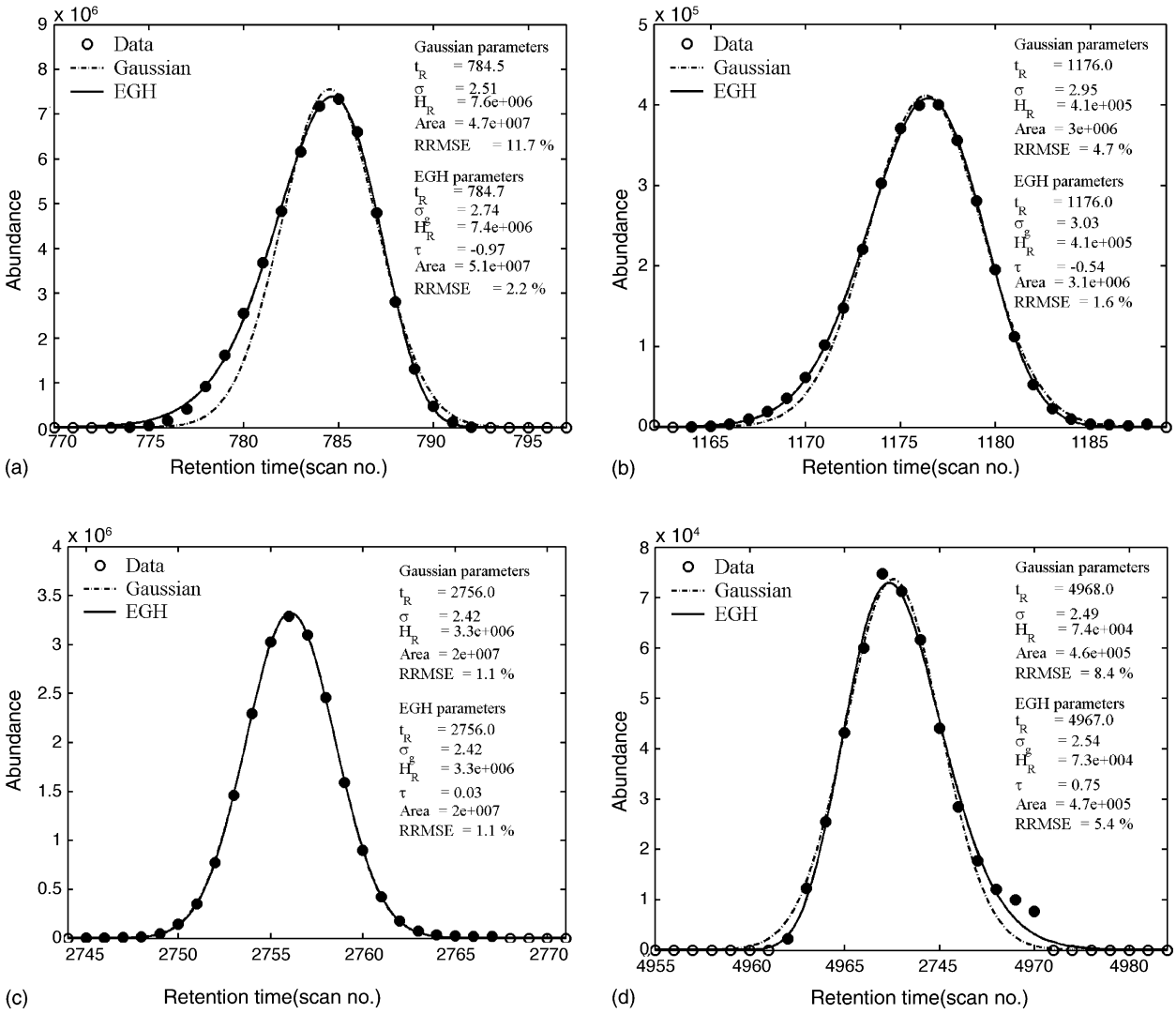

Fig. 5. Gaussian and EGH fit of chromatographic peaks. (a) Naphthalene (moderately fronting, $\tau = -0.97$), C_1-benzothiophene isomer (fronting, $\tau = -0.54$), phenanthrene (symmetrical, $\tau = 0.03$), 13β(H), 17α(H), 20R-diacholestane (moderately tailing, $\tau = 0.75$). Peak parameters (t_R, σ, H_R, and τ), peak areas, and RRMSE are listed in plots. Data points used for calculating RRMSE are shown as filled circles.

second peak) from well-resolved peaks, resolved but baseline distorted, and for peak clusters of incompletely resolved peaks. Fig. 7a and b show the relative standard deviation (R.S.D.) (as defined in [34]) for 25, 37, and 10 DRs, using peak areas or heights, respectively.

For well-resolved peaks the Gaussian and EGH peak functions based on peak areas gave comparable low uncertainties with R.S.D.s < 3.2%. Conversely, for baseline distorted peaks the R.S.D.s were lower for the Gaussian function (1.5–6%) than the EGH (1.5–16%). The large R.S.D. of some peaks were caused by data points affected by distorted baselines, and reducing the number of data points used in the parameter estimation to $\pm\sigma$ solved the problem. Accordingly, we recommend leaving out the most heavily affected data points if a modified Gaussian peak function is used for approximating baseline distorted peaks. The use of peak heights (Fig. 7b) reduced the uncertainties to <2.2%, and <3.5% for well resolved and baseline distorted peaks, respectively, using the Gaussian function, and to <2.2%, and <8%, for EGH.

Furthermore, results indicate that the precisions of DRs, calculated on the basis of Gaussian peak estimates, are comparable with those obtained with commercial software. For well-resolved peaks, R.S.D.s were <3.2 (areas) and <2.2% (heights), compared to <2.5 and <2.8%, respectively, with commercial software. For baseline distorted peaks R.S.D.s were <5.7 and <3.5% compared to <4.6 and <4.0%, and finally for incompletely resolved peaks R.S.D.s were <7.2 and <6.5% using Gaussian peak estimates, compared to <7.9 and 5%, with commercial software.

Although, the data quality (based on R.S.D.s) of data obtained using our approach is comparable with commercial software, mean values of some DRs depend on the method applied. For fronting and tailing peaks the Gaussian peak estimates (areas and heights) deviate from those

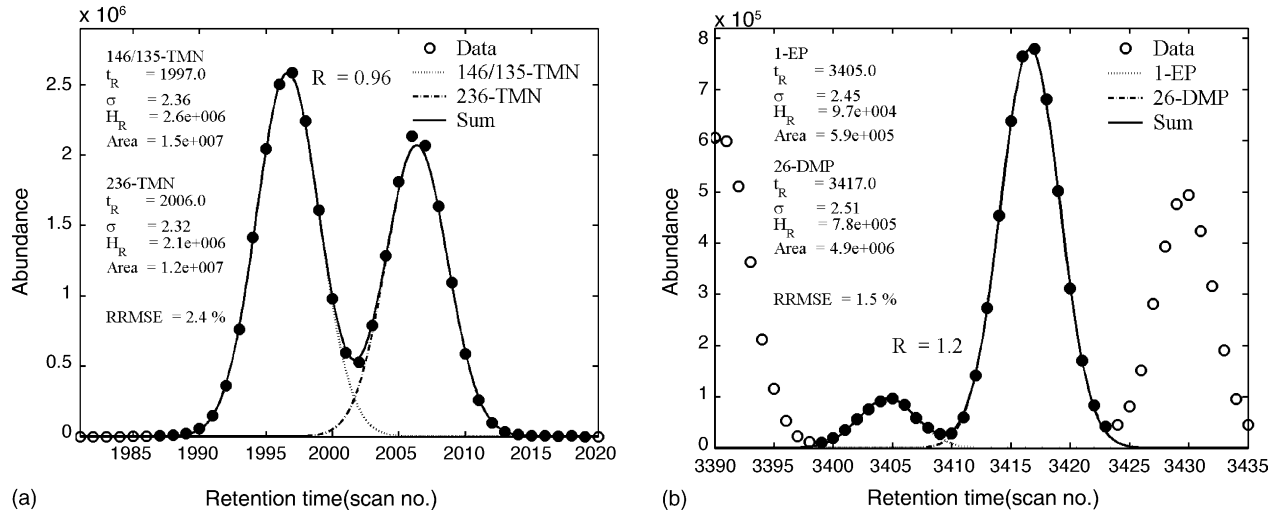

Fig. 6. Gaussian fits of incompletely resolved chromatographic peaks. (a) C_3-naphthalenes, 1,4,6-/1,3,5-trimethylnaphthalene (146-/135-TMN) and 2,3,6-trimethylnaphthalene (236-TMN) with height ratio of approximately 1:1; (b) C_2-phenanthrenes, 1-ethylphenanthrene (1-EP) and 2,6-dimethylphenanthrene (26-DMP) with height ratio of approximately 1:8. Peak parameters (t_R, σ, and H_R), peak areas, RRMSE, and chromatographic resolution (R) are listed in plots. Data points used for calculating RRMSE are shown as filled circles.

obtained with the two other methods (data not shown). For baseline distorted and especially incompletely resolved peaks, however, data obtained with commercial software deviate systematically from those obtained with the two peak functions.

Generally, DRs based on peak heights were less uncertain compared to those based on areas. This seems reasonable since baseline distortion and the properties of the applied peak function affect peak height to a lesser extent than they affect peak areas. DRs calculated for analytical samples can be normalized to the corresponding DRs in the laboratory reference sample analyzed closest in time [31]. Hence, the effects of changes in peak shape, which affects heights more than peak areas, can be reduced. Accordingly, we recommend peak estimations using the Gaussian function followed by calculations of DRs from peak heights (H_R) as the optimal data preprocessing procedure.

Before a peak can be detected in our approach (peak detection limit) there need to be a data point within the peak region, which exceed the noise multiplied with d_{min}, and the number of data points within this region must allow initial parameterization of t_R, H_R, and σ. Furthermore, replicate laboratory reference samples may be used for quality control of data (control limits). We have previously applied our data preprocessing procedure to GC–MS chromatographic data prior to PCA. Peak areas based on the Gaussian function were used to calculate 88 DRs of PACs and petroleum biomarkers for use in forensic oil spill identification [31]. DRs were normalized to the laboratory reference oil, which resulted in low analytical standard deviations (between 0.05 and 3.2%), comparable

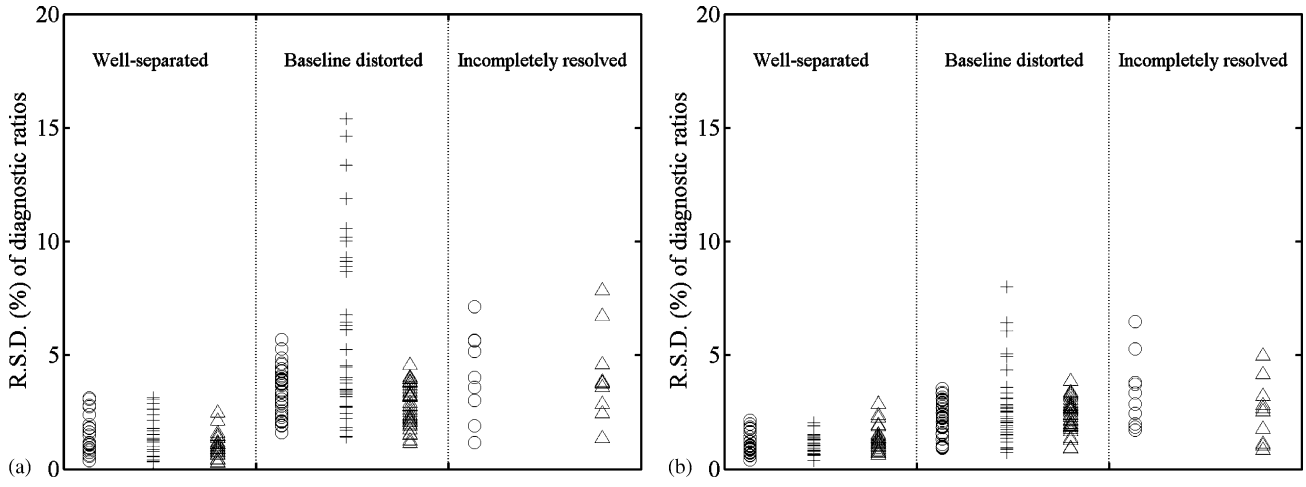

Fig. 7. Precision of DRs using Gaussian, EGH, and commercial software. (a) R.S.D.s of 25 DRs comprised of peak areas of well-resolved peaks, 37 of baseline distorted, and 10 comprised of incompletely resolved peaks. (b) DRs comprised of peak heights. Symbols (◯), (+), and (△) denote R.S.D.s calculated with Gaussian, EGH, and commercial software, respectively.

to or smaller than those, obtained by a quantitative method. Furthermore, the integrated methodology presented in [31] comprised among others a selection criterion for identifying DRs with large R.S.D.s. Thus, if highly uncertain DRs are present in data, e.g. introduced by the data preprocessing, these ratios can be detected and subsequently deselected prior to analysis.

4. Conclusion

A new data preprocessing method specific for processing first-order data from hyphenated analytical techniques (e.g. GC–MS) was developed and validated in this article. The method combines several procedures for data preprocessing and enables semi-automatic transformation of chromatograms into semi-quantitative variables (e.g. DRs) and their uncertainties, which are directly accessible for chemometric data analysis. Compared to commercial software, the method presented here is less time consuming and more objective. Furthermore, it handles retention time shifts, which can be a large impediment for standard chromatographic processing. Eventually, it presents an advantage over commercial software for resolving overlapping peaks. The perpendicular-drop or tangent-skimming algorithms incorporated in commercial integrators can introduce systematic errors in peak area and height calculations. Hence, our method enables fast screening analysis of complex chemical mixtures for e.g. environmental monitoring and forensics, that otherwise would be cumbersome and time-consuming and thus practically inapplicable. The suggested preprocessing method was validated using analytical data from 120 oil components belonging to several chemical classes. DRs based on peak areas and heights were calculated from well-resolved, baseline distorted and incompletely resolved peaks in 20 replicate laboratory reference oil samples. In addition, the method has been applied for data preprocessing and calculation of diagnostic fingerprinting ratios in relation to forensic oil spill identification.

Acknowledgements

This work was funded by the Department of Environmental Chemistry and Microbiology, National Environmental Research Institute, and by the Natural Sciences Research Council, Denmark.

References

[1] M. Frignani, L.G. Bellucci, M. Favotto, S. Albertazzi, Hydrobiologia 494 (2003) 283.
[2] I. Johansson, B. van Bavel, Sci. Total Environ. 311 (2003) 221.
[3] A. Stella, M.T. Piccardo, R. Coradeghini, A. Redaelli, S. Lanteri, C. Armanino, F. Valerio, Anal. Chim. Acta 461 (2002) 201.
[4] Z.D. Wang, M. Fingas, D.S. Page, J. Chromatogr. A 843 (1999) 369.
[5] G. De Luca, A. Furesi, R. Leardi, G. Micera, A. Panzanelli, P.C. Piu, G. Sanna, Mar. Chem. 86 (2004) 15.
[6] W.A. Burns, P.J. Mankiewicz, A.E. Bence, D.S. Page, K.R. Parker, Environ. Toxicol. Chem. 16 (1997) 1119.
[7] S.A. Stout, V.S. Magar, R.M. Uhler, J. Ickes, J. Abbott, R. Brenner, Environ. Forensics 2 (2001) 287.
[8] R. Bro, Anal. Chim. Acta 500 (2003) 185.
[9] M. Johansson, M. Berglund, D.C. Baxter, Spectrochim. Acta Part B: Atomic Spectrosc. 48 (1993) 1393.
[10] V.R. Meyer, J. Chromatogr. Sci. 33 (1995) 26.
[11] N. Dyson, J. Chromatogr. A 842 (1999) 321.
[12] J.H. Christensen, G. Tomasi, A.B. Hansen, Environ. Sci. Technol., in press.
[13] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.
[14] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71.
[15] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[16] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154.
[17] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649.
[18] P.H.C. Eilers, Anal. Chem. 76 (2004) 404.
[19] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.
[20] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231.
[21] I. Eide, G. Neverdal, B. Thorvaldsen, H.L. Shen, B. Grung, O. Kvalheim, Environ. Sci. Technol. 35 (2001) 2314.
[22] B.V. Grande, R. Manne, Chemometr. Intell. Lab. Syst. 50 (2000) 19.
[23] R. Manne, B.V. Grande, Chemometr. Intell. Lab. Syst. 50 (2000) 35.
[24] F. Gong, Y.Z. Liang, Y.S. Fung, F.T. Chau, J. Chromatogr. A 1029 (2004) 173.
[25] H.L. Shen, R. Manne, Q.S. Xu, D.Z. Chen, Y.Z. Liang, Chemometr. Intell. Lab. Syst. 45 (1999) 323.
[26] V.B. Di Marco, G.G. Bombi, J. Chromatogr. A 931 (2001) 1.
[27] S. Levent, Anal. Chim. Acta 312 (1995) 263.
[28] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou, J. Chromatogr. A 912 (2001) 13.
[29] J.R. Torres-Lapasio, J.J. Baeza-Baeza, M.C. Garcia-Alvarez-Coque, Anal. Chem. 69 (1997) 3822.
[30] P.D. Boehm, G.S. Douglas, W.A. Burns, P.J. Mankiewicz, D.S. Page, A.E. Bence, Mar. Pollut. Bull. 34 (1997) 599.
[31] J.H. Christensen, A.B. Hansen, J. Mortensen, G. Tomasi, O. Andersen, Environ. Sci. Technol. 38 (2004) 2912.
[32] P.S. Daling, L.G. Faksness, A.B. Hansen, S.A. Stout, Environ. Forensics 3 (2002) 263–278.
[33] S.A. Stout, A.D. Uhler, K.J. McCarthy, Environ. Forensics 2 (2001) 87.
[34] J.C. Miller, J.N. Miller, Statistics for Analytical Chemistry, Ellis Horwood and Prentice Hall, Hartnolls, Bodmin, Great Britain, 1993.
[35] S.C. Pai, J. Chromatogr. A 988 (2003) 233.
[36] M.S. Jeansonne, J.P. Foley, J. Chromatogr. Sci. 29 (1991) 258.
[37] J.W. Li, J. Chromatogr. A 952 (2002) 63.
[38] K. Lan, J.W. Jorgenson, J. Chromatogr. A 915 (2001) 1.
[39] J.A. Nelder, R. Mead, Comput. J. 7 (1965) 308.
[40] A.O. Barakat, A.R. Mostafa, J. Rullkotter, A.R. Hegazi, Mar. Pollut. Bull. 38 (1999) 535.