Talanta 81 (2010) 120–128

Contents lists available at ScienceDirect

Talanta

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

Chemometric analysis of gas chromatography–mass spectrometry data using fast retention time alignment via a total ion current shift function

Jeremy S. Nadeau[a], Bob W. Wright[b], Robert E. Synovec[a,*]

[a] Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA
[b] Pacific Northwest National Laboratory, Battelle Boulevard, P.O. Box 999, Richland, WA 99352, USA

ARTICLE INFO

Article history:
Received 21 September 2009
Received in revised form
16 November 2009
Accepted 17 November 2009
Available online 4 December 2009

Keywords:
Alignment
Gas chromatography
Mass spectrometry
Principal component analysis

ABSTRACT

A critical comparison of methods for correcting severely retention time shifted gas chromatography–mass spectrometry (GC–MS) data is presented. The method reported herein is an adaptation to the piecewise alignment algorithm to quickly align severely shifted one-dimensional (1D) total ion current (TIC) data, then applying these shifts to broadly align all mass channels throughout the separation, referred to as a TIC shift function (SF). The maximum shift varied from (−) 5 s in the beginning of the chromatographic separation to (+) 20 s toward the end of the separation, equivalent to a maximum shift of over 5 peak widths. Implementing the TIC shift function (TIC SF) prior to Fisher Ratio (*F*-Ratio) feature selection and then principal component analysis (PCA) was found to be a viable approach to classify complex chromatograms, that in this study were obtained from GC–MS separations of three gasoline samples serving as complex test mixtures, referred to as types C, M and S. The reported alignment algorithm via the TIC SF approach corrects for large dynamic shifting in the data as well as subtle peak-to-peak shifts. The benefits of the overall TIC SF alignment and feature selection approach were quantified using the degree-of-class separation (DCS) metric of the PCA scores plots using the type C and M samples, since they were the most similar, and thus the most challenging samples to properly classify. The DCS values showed an increase from an initial value of essentially zero for the unaligned GC-TIC data to a value of 7.9 following alignment; however, the DCS was unchanged by feature selection using *F*-Ratios for the GC-TIC data. The full mass spectral data provided an increase to a final DCS of 13.7 after alignment and two-dimensional (2D) *F*-Ratio feature selection.

1. Introduction

Gas chromatography (GC) is an important chemical analysis tool for separating complex mixtures [1–3]. However, for many GC applications for truly complex samples, the analyst is faced with the reality that it is not practical, and indeed not possible, to physically resolve all of the compounds of interest [4–10]. Additionally, traditional data analysis strategies such as compound identification through "peak" retention time and mass spectral matching with standards followed by integration for quantification purposes are not as appealing if many of the important analyte compounds are overlapped with interfering compounds; if these traditional data analysis strategies are applied, the results are often unacceptable. In these situations, multivariate "chemometric" data analysis methods are often relied upon to mathematically complete the analysis [8–19].

Many studies have shown how powerful initial pre-processing, like retention time alignment, can be to glean information from chromatographic data [4,8,10,20–26]. This pre-processing correction of the GC data must maintain the desired chemical information [22,24,25,27]. Indeed, retention time precision has been clearly shown to be an important component in many types of semi-automated and automated analyses [15,22,24,26,28,29]. Obtaining good results from chemometric methods such as the generalized rank annihilation method (GRAM), principal component analysis (PCA) and partial least squares (PLS) all share this need for improving retention time precision via an objective alignment prior to applying these chemometric algorithms [3,4,6,8,9,11,12,14,17,18,20,22,30,31].

Satisfactory retention time precision prior to applying chemometric algorithms is essential no matter what the order of the data is, i.e., first order as that produced by GC coupled with flame ionization detection (GC-FID) [4,9,19,21,22,24,25,32], second order such as that produced by GC coupled with mass spectrometry (GC–MS) [3,6,8,10,11,13–15,20,23,28,33,34], and even in third order data, such as that produced by comprehensive two-dimensional GC coupled with time-of-flight mass spectrometry (GC × GC-TOFMS)

* Corresponding author.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).

0039-9140/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2009.11.046

[35,36]. Even with the extra chemical information afforded by the other dimensions, analysis of higher order GC data generally benefits by improving the retention time precision through an alignment pre-processing procedure.

Many algorithms have been developed to obtain the retention time precision necessary for chemometric analysis. The major methods for this purpose include peak matching [28,32], windowed shifting [9,10,22,31], correlation optimized warping (COW) [13,21,23–25,37], and rank minimization [16,29,30,38,39]. Rank minimization is specifically tailored to the alignment of second order data, e.g., GC–MS or GC × GC-FID. However, rank minimization (and peak matching algorithms in general) is useful mainly for data that are shifted less than one chromatographic peak width. This limits or even prevents their usefulness with severely shifted data. However, the windowed shifting and COW algorithms are very useful for severely shifted data [9,21–25,37]. Unfortunately, there are few methods that combine the speed and robustness of the windowed shifting or COW algorithms with the precision of the rank minimization or other peak matching algorithms. Methods that do incorporate the alignment of multi-dimensional data with methods like COW still require a correlation calculation of the segments with a larger amount of data, thus slowing their performance for severely shifted data [21,23]. Initial alignment is one way to both incorporate the robust algorithms and retain the precise alignment that the multi-dimensional data allows [23]. However, coarse shifting of the entire data set, if not properly implemented, can incorrectly align segments of the data.

Another aspect of various alignment algorithms that can be important is the ability, or lack thereof, to view the path the alignment algorithm took to improve the retention time precision [23,32–34,40]. Most of the methods provide a procedure for correcting the shifting but fewer offer visual inspection of the shifts. An example of the benefit of shift inspections can be seen with the XCMS algorithm, which first uses a peak finding algorithm to find the similar peaks and shifts all the peaks into alignment, and after a second fine tuned alignment, it calculates retention time deviations, which are used to align the mass spectral information [40]. These shift deviations could be quite helpful in comparing the precision of a set of chromatographic separations or even the change of column condition over time. The idea implemented in this report is to use this information in a diagnostic format. If the run-to-run shift deviations are put in vector format, and defined along the chromatographic time axis, they comprise a "shift function," designated SF, that can be used to align one chromatogram to another. A key benefit of using a SF is to speed up the alignment of the full GC–MS data, which is demonstrated herein, retaining the spectral information, in order to facilitate second order chemometric data analysis.

In this report, a method of alignment that utilizes a SF based upon the total ion current (TIC) of GC–MS data is developed and explored. The basic idea is to align the TIC chromatograms for a set of GC–MS data, and then to keep track of the TIC SF for each GC–MS chromatogram relative to a chosen common target chromatogram. The TIC SF incorporates both a coarse alignment (to correct large shifting) and secondary sub-alignment (to correct subtle peak-to-peak shifting) steps. The subsequent TIC SF for each GC–MS chromatogram is then applied across the board to all m/z chromatograms. In this way, all of the *entire* GC–MS chromatograms can be aligned at all m/z, thus preserving the integrity of the second order data structure. A goal is to demonstrate the ability of the TIC SF approach to improve the quality of the data prior to chemometric data analysis, specifically PCA in this report. The ultimate goal of the TIC SF-based algorithm is to be able to quickly align complex GC–MS data for the purpose of optimal chemometric data analysis. The method reported herein finds the shifts of every point in the TIC, with a piecewise alignment algorithm [9,10,22], and then using this SF, aligns all the m/z ion chromatograms along the chromatographic time dimension. The method has advantages in applicability and memory usage. Additionally, and very importantly, the TIC SF can be successfully applied to the chromatograms at specific m/z channels that do not have enough peaks by themselves to independently properly align. The notion is that the TIC represents the "most complex version" of a given sample, and thus the TIC SF will provide the most robust alignment of all m/z relative to each other within a given GC–MS chromatogram and from one chromatogram to another.

To test the performance of the TIC SF alignment approach, two sufficiently different temperature programs are used with three gasoline samples, as test mixtures, with chromatograms collected using GC–MS. The retention time shifting that will be corrected by the TIC SF approach is significantly larger (several peak widths and variable) than that corrected in a prior study (a fraction of the peak width) [10]. The two programs introduce substantial run-to-run retention time shifting, that is akin to that confronted when researchers need to analyze chromatographic data sets that have been collected over long periods of time such as in an industrial application. The two temperature programs produce very challenging data to align across all m/z for the entire length of the chromatograms. Principal component analysis (PCA) is used to examine the performance of the TIC SF-based alignment for the application of classifying samples, since PCA performance is very sensitive to having the data properly registered from sample to sample. In this case, PCA of the TIC chromatograms (as a benchmark of first order GC-TIC data) is compared to PCA of the entire GC–MS data that had each m/z aligned using the TIC SF methodology. Subsequently, a comparison is also made using specific m/z that have been selected via an ANOVA based Fisher Ratio (F-Ratio) approach using a training set of two complex samples. For these studies, gasoline samples are used as model complex samples.

2. Experimental

All GC separations were performed on an Agilent 6890GC equipped with a 5973A Mass Spectrometer with unit mass resolution, and a 7683B auto injector, equipped with electronic pressure control (Agilent, Palo Alto, CA, USA). Unleaded gasoline samples were collected from fueling stations as previously described [9]. The gasoline samples were arbitrarily labeled Type C (C), Type M (M), and Type S (S). All gasoline samples were separated on a 30.0 m DB-5 fused silica column with 0.25 mm i.d. and 0.2 μm DB-5 (Agilent, Palo Alto, CA, USA) stationary phase. The inlet temperature was 275 °C. Two temperature programs were used to induce severe shifting to purposely challenge the TIC SF algorithm. Temperature program 1 (P1) was initiated at 40 °C for 1 min then increased at a rate of 10 °C/min to 80 °C, then increased at a rate of 20 °C/min to 240 °C and held constant for 2 min. Temperature program 2 (P2) was initiated at 40 °C for 0.85 min then increased at a rate of 10 °C/min to 80 °C and held for 0.5 min then increased at a rate of 20 °C/min to 240 °C and held for 1.65 min. The flow rate for both programs was held constant at 1.5 ml/min with an initial head pressure of 3.25 psi (22.4 kPa) using H_2 as the carrier gas.

The GC–MS data provided by the Chemstation software were extracted and imported into Matlab 7.0.4 (The Mathworks, Inc., Natick, MA, USA) for further manipulation. All chromatograms were initially unskewed to remove the concentration bias caused by the data collection of each ion not being sufficiently faster than the chromatographic separation, i.e., to keep up with the eluting peak concentration changing sufficiently quickly in relation to the mass spectrum scanning rate [15,41]. All chromatograms were baseline corrected by taking the first and last 16 s of each ion chromatogram and fitting a straight line through the baseline to remove either an increasing slope or decreasing slope, and any offset. The data were

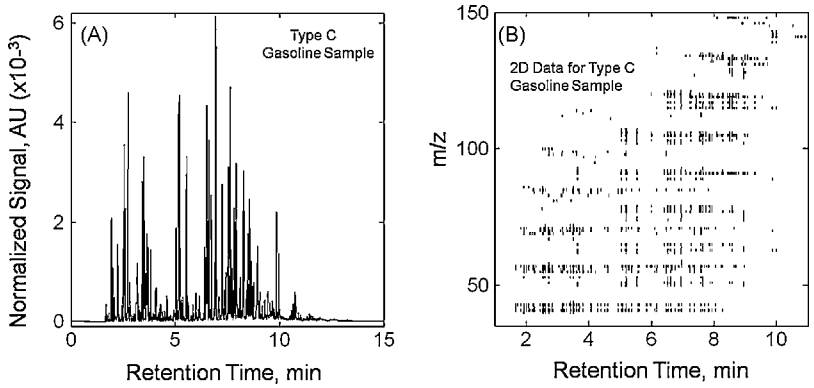

Fig. 1. (A) Typical chromatogram of a gasoline sample using temperature program P1, with conditions provided in Section 2. (B) The full mass spectral 2D plot of the separation from (A).

then normalized to the total signal of all the ion chromatograms for a given sample. Subsequently, the data were analyzed using PCA software (PLS_Toolbox Version 4.2.1, Eigenvector Research, Inc., Wenatchee, WA, USA).

3. Theory

The alignment algorithm implemented in this report using the TIC SF is intrinsically based upon an algorithm reported by Pierce and co-workers [9,10,22]. The steps in the alignment follow the general form previously described [9,32]. In the first step, the signal for all the ions at each retention time for a given GC run are added together to produce the total ion chromatogram (TIC). The TIC is fractionated into windows (W), smaller data sections from the overall separation of equal length, and the windows are iteratively shifted, limited by a maximum shift value (L) to match to an arbitrarily chosen target TIC. The target TIC is one of the samples chosen as a retention time reference for all the other samples, including the target, to be aligned to it. The shift of each of the windows is the shift that gives the maximum correlation coefficient between the sample and the target [9]. After all of the shifts of the TIC for each of the windows are calculated, a function of the new locations for every data point is linearly interpolated. The resulting function is called the TIC SF for a given GC–MS chromatogram. The TIC SF is used to interpolate between points at every m/z ion for the given GC–MS chromatogram. At this point, a ratio of the number of points

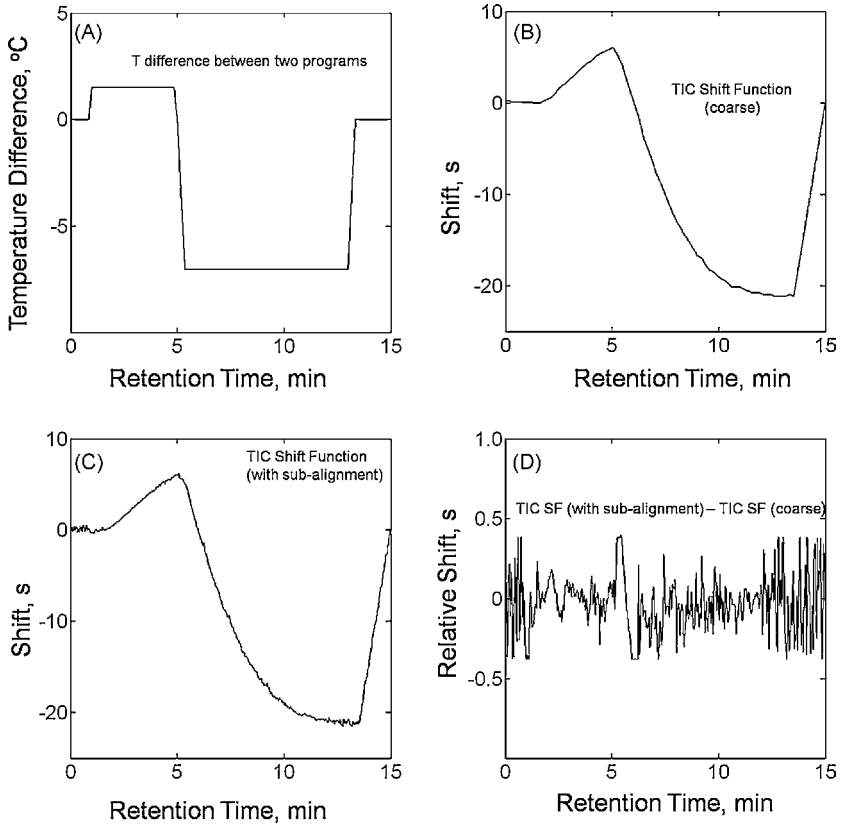

Fig. 2. (A) Difference in the temperature between the two programs is plotted as a function of time: P2 – P1. (B) The total ion current shift function (TIC SF), obtained as explained in Section 3, shows that the different hold times cause the peaks in temperature program P2 to come out about 10 s faster in the initial portion of the chromatogram and lag behind by about 20 s at the end of the chromatogram. (C) The TIC SF following sub-alignment, corrects the subtle shifting in the chromatogram (as explained in Section 3). (D) Subtraction of the TIC SF with sub-alignment from the TIC SF (coarse) without sub-alignment: (C) – (B). It is possible to see all the locations where a second stage of alignment provides fine tuning of the retention time alignment.

in the window before and after interpolation is produced. This ratio is multiplied by the window to ensure the conservation of peak area for each window after alignment. A secondary alignment step was applied to correct the more subtle shifting that remained after the initial "coarse" alignment. In this secondary alignment step, the coarse aligned data matrix, $m \times n$, and target, $1 \times n$, are linearly interpolated to make matrices which are $5n$ points long. Hence, the window size and shift parameters in the alignment software were reduced 5-fold to provide a refined, and hence more precise, alignment, i.e., referred to as sub-alignment. This step in the alignment corrects for the subtle peak-to-peak shifting seen in the data and can correct for shifting that is smaller than the collection rate of the instrument. The resulting TIC SF from the secondary alignment step is then combined with the coarse TIC SF to produce an overall TIC SF used to align the original unaligned GC–MS data with n time points in the separation. The TIC SF approach essentially uses the "most complex version" of the sample, thus avoids trying (unsuccessfully) to align all m/z individually, which is not feasible since at many m/z there is not enough signal relative to the baseline to provide a reliable individual m/z SF. Thus, the TIC SF is a global approach to correct misalignment for all m/z, presuming there is not too much chemical selectivity-based shifting from one m/z to the next. The issue of chemical selectivity-based shifting (e.g., from run-to-run column degradation) can be addressed by testing the TIC SF occasionally to be "under control."

4. Results and discussion

4.1. The total ion current (TIC) shift function (SF)

Before alignment was performed, six replicates of three gasoline types (C, M and S) were run on two different temperature programs, P1 and P2 (36 chromatograms total). Note that there are two primary sources of retention time shifting: the large shifting due to the different temperature programs and high frequency shifting (misalignment "noise") due to uncontrollable subtle flow rate and temperature fluctuations inherent to running the GC instrument, even though the instrument was run under electronic pressure control. The original (raw, unaligned) gasoline chromatogram of the Type C gasoline sample is shown in Fig. 1A, representing the summation of the mass spectral information (TIC), of the full GC–MS chromatogram shown in Fig. 1B.

The temperature difference between the two temperature programs, P2 – P1, is shown in Fig. 2A. The program P1 was used for the target chromatogram for alignment essentially making it the "target program." A TIC Shift Function (TIC SF) in Fig. 2B is the point-by-point correction of the retention times for the components in the sample chromatogram (to be aligned) minus the retention times for the components in the target chromatogram. The separation from temperature program P2 lags behind the target by about 5 s in the initial segment of the separation and pushes beyond the target by about 20 s in the latter part of the separation. The difference in the two temperature programs severely shifted the data beyond more than ~5 neighboring peaks, which gives a maximum shift observed from this cause of ~22 s. The TIC SF in Fig. 2B also provides the retention time shifts necessary to correct the misalignment for the chromatograms at all m/z in the one chromatogram relative to the other, presuming sufficient consistency in the chemical selectivity provided by the chromatography, as will be demonstrated. The TIC SF pictorially shows how the two chromatographic separations differ under the two temperature programs, and also shows the path the algorithm must take to correctly align all m/z. As seen in Fig. 2B, the shifts of the data are dynamic. For the piecewise alignment algorithm, large and dynamic shifting requires a large window for correct alignment of these data. Because of this large window, only the major shifts are initially corrected, and the subtle

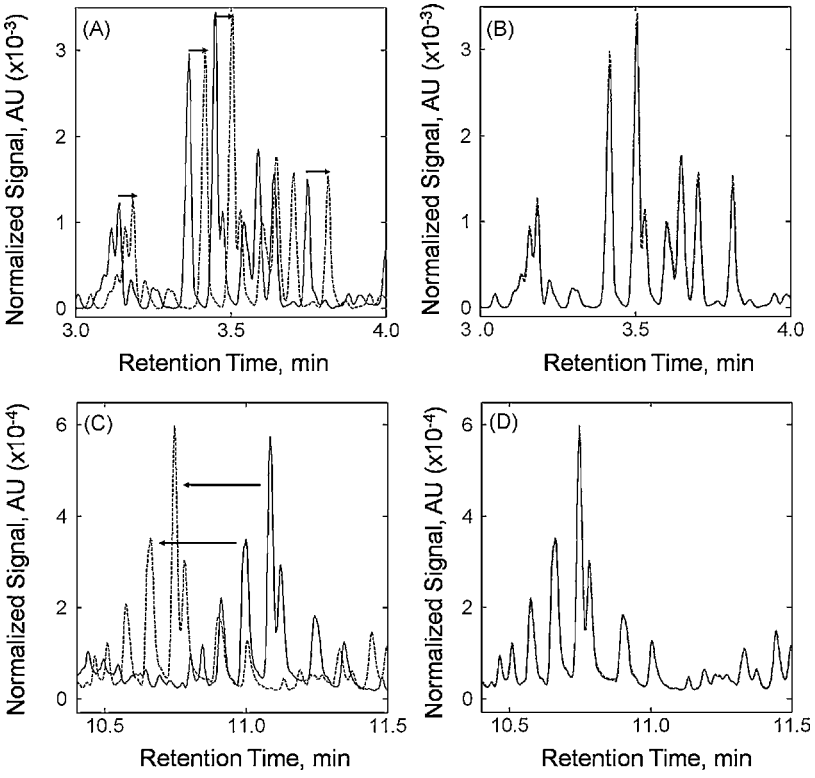

Fig. 3. (A) Run-to-run retention time shifting in a portion of the data in the beginning of the chromatogram is illustrated by this zoom-in of Fig. 1. (B) Same region as in (A), following correction using the TIC SF with sub-alignment (i.e., in Fig. 2C). (C) Run-to-run retention time shifting in a portion of the data toward the end of the chromatogram is illustrated by this zoom-in of Fig. 1. Note that shifting is in opposite direction of that shown in (A). (D) Same region as in (C), following correction using the TIC SF with sub-alignment (i.e., in Fig. 2C).

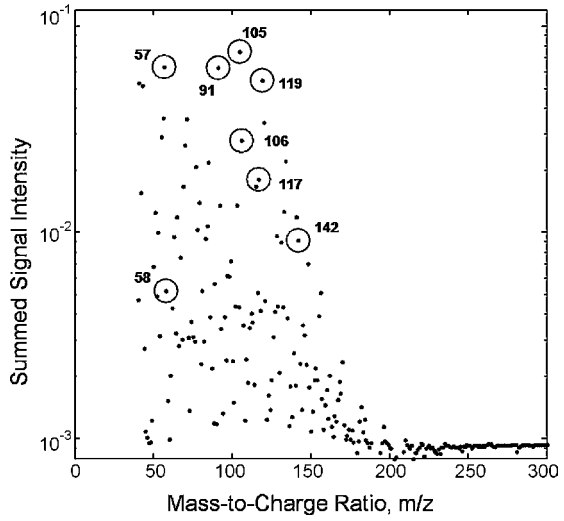

Fig. 4. Ions were identified to compare the individual *m*/*z* ion SFs to the TIC SF. In this plot, the largest signals were plotted to find ions that were selective for different functional groups, and/or had a substantial total signal that would in principle lead to an improved alignment of chromatograms.

peak-to-peak shifts are not corrected for until a sub-alignment is applied where the window size represents the approximate peak width of the data. The TIC SF with sub-alignment (Fig. 2C) from this secondary alignment step corrects the subtle peak-to-peak shifting previously uncorrected. Because the sub-alignment step corrects for the more subtle "high frequency" shifting, the TIC SF with sub-alignment differs slightly from the coarse TIC SF. The difference between the two functions (i.e., subtracting the TIC SF with coarse alignment only from TIC SF with sub-alignment) is visible in Fig. 2D. The differences between these SFs are never more than 0.5 s and are generally centered around zero.

We turn our attention now to describe how the TIC SF incorporating both coarse and sub-alignment as in Fig. 2C performs when implemented. For example, a sample data run on P2 lag behind the target data on P1 in the region between 3 and 4 min (Fig. 3A). The two separations are superimposed after applying the TIC SF with sub-alignment (Fig. 3B), indicating the negative shifts at the beginning of the chromatogram are eliminated. In a similar manner, Fig. 3C shows the large positive shifting of the sample relative to the target near the end of the chromatogram. Again, in Fig. 3D, the sample and target are essentially superimposed on each other after alignment.

4.2. Using individual m/z ion shift functions to test the TIC shift function for robustness

The alignment of all *m*/*z* ions using the TIC SF-based alignment algorithm relies upon the condition that the TIC SF can sufficiently explain (i.e., correct for) the retention time imprecision for each and every *m*/*z* ion chromatogram in the GC–MS data. To demonstrate how the shift of a given TIC chromatogram run on temperature program P2 compares to the actual shifting of the ions in these data, 8 representative ions (*m*/*z* 57 and 58 for alkane groups, *m*/*z* 91, 105, 106, 119, and 117 aromatic groups and *m*/*z* 142 for low signal-to-noise) with variable signal were aligned to a target chromatogram from temperature program P1. All the *m*/*z* ions are shown in Fig. 4 with the selected ions circled and labeled for reference. The window size and maximum shift were optimized for each of the ions and with the optimum W and L values each ion was aligned to the target ion from temperature program P1. The alignment again produced shift functions at the individual *m*/*z*.

An example of a well-behaved ion is shown with *m*/*z* 57, and an example of an ill behaved ion is shown with *m*/*z* 106. Ions with higher signal aligned better in general, but based on signal, no general trend could be drawn in this regard. The *m*/*z* SFs and TIC SFs both with sub-alignment are shown relative to the coarse TIC SF, analogous to how the TIC SF (with sub-alignment) was shown relative to the TIC SF (coarse alignment only) in Fig. 2D. However, the individual ion *m*/*z* SFs with sub-alignment are only shown in regions where there are peaks present. The individual *m*/*z* 57 SF with sub-alignment in Fig. 5A behaves in a very similar manner to the TIC SF with sub-alignment. As shown in Fig. 5B, in the latter region of the separation the sample peaks are shifted positively relative to the target peaks prior to alignment. The alignment of *m*/*z* 57 using the *m*/*z* 57 SF with sub-alignment is shown in Fig. 5C, and using the TIC SF with sub-alignment is shown in Fig. 5D. Both the *m*/*z* 57 SF with sub-alignment and the TIC SF with sub-alignment corrected the retention time imprecision observed in Fig. 5B. Another way to evaluate the benefits (or lack thereof) of implementing a given SF is to show important regions indicated by a Fisher Ratio (*F*-Ratio) analysis using a training set with closely related samples such as the type C and M samples in this study. The *F*-Ratios in this context would find retention time locations in which the peaks in the aligned data were, with statistical significance, different in the type C and M samples. The *F*-Ratios using the type C and M samples are shown in Fig. 5E after *m*/*z* 57 SF alignment (lower plot) and after TIC SF alignment (upper plot). Results from the *F*-Ratio analysis on the *m*/*z* 57 mass channel data show that the same two major regions indicated as being significantly different with the *m*/*z* 57 SF and the TIC SF, with the *F*-Ratios for the TIC SF aligned data appearing to be slightly more sensitive (e.g., the peak at 7.8 min with an *F*-Ratio of about 400).

In contrast to the results for *m*/*z* 57 are the results for *m*/*z* 106. The TIC SF subtracted from the *m*/*z* 106 SF and the TIC SF both with sub-alignment are shown in Fig. 6A. The *m*/*z* 106 SF with sub-alignment differed significantly from the TIC SF with sub-alignment. There is a significant negative spike in the *m*/*z* 106 SF plot, indicating improper alignment. The actual *m*/*z* 106 data shown in Fig. 6B indicates the shifting of the sample peaks relative to the target peaks was again positive. Alignment using the *m*/*z* 106 SF with sub-alignment is shown in Fig. 6C while alignment using the TIC SF with sub-alignment is shown in Fig. 6D. When using the *m*/*z* 106 SF to correct the misalignment, several peaks were improperly corrected, but using the TIC SF, the data were correctly aligned. The *F*-Ratios between the type C and type M samples after TIC SF with sub-alignment and *m*/*z* 106 SF with sub-alignment are shown in Fig. 6E. Note that the two *F*-Ratio plots are on very different scales since the *F*-Ratios for the *m*/*z* 106 SF aligned data were very small due to significant misalignment throughout the entire type C and M training set. The results show that *F*-Ratios obtained after the individual *m*/*z* 106 alignment is essentially just noise and no useful information can be gleaned from this analysis. However, the TIC SF alignment provided *F*-Ratios that could be used to find regions of interest, i.e., peaks that change with statistical significance between the type C and M samples such as the peak at 8.1 min with an *F*-Ratio of nearly 400. For the *m*/*z* ions evaluated, *m*/*z* 57, 91, 105, 119, 58, and 142 all behaved well using either the TIC SF with sub-alignment or the individual *m*/*z* SFs with sub-alignment. However, *m*/*z* 106 and 117 were corrected only by using the TIC SF with sub-alignment, and the individual *m*/*z* SFs were problematic for these. In general, for many of the *m*/*z* ions, as indicated in Fig. 4, the individual *m*/*z* SFs would be very problematic due to low signal-to-noise (essentially absence of enough peaks at the given *m*/*z*). The use of the TIC SF makes it possible to align all *m*/*z* without any consideration to signal or peak locations in the individual *m*/*z* data. The improved accuracy using the TIC SF is evidence that using the TIC SF is a reasonable approximation for describing the retention time

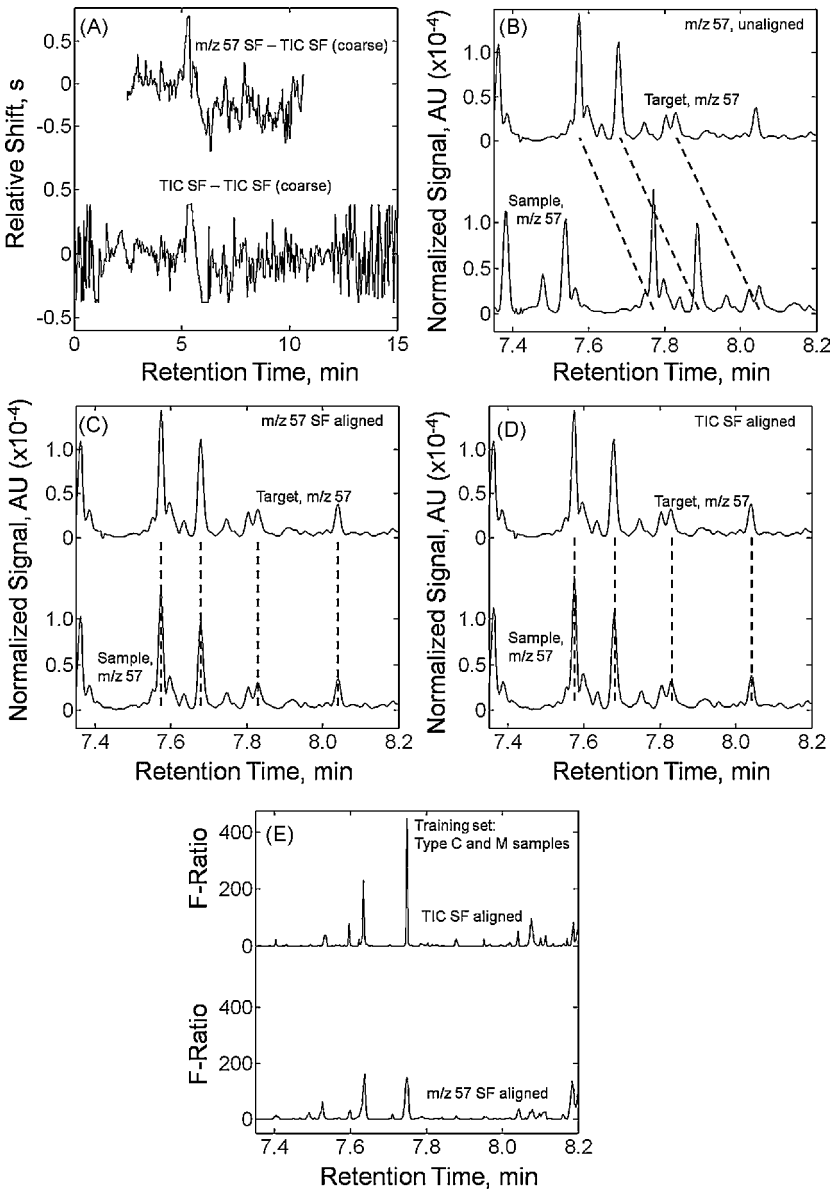

Fig. 5. (A) The SF (with sub-alignment) for *m*/*z* ion 57, and the TIC SF (with sub-alignment), both relative to the TIC SF (coarse, without sub-alignment) show similar behavior over the entire signal range. See Fig. 2D for example of plot. (B) Data portion illustrating misalignment of the *m*/*z* 57 ion. (C) Misalignment as in (B) is corrected by applying the *m*/*z* 57 SF with sub-alignment. (D) Misalignment as in (B) is corrected by applying the TIC SF with sub-alignment. (E) *F*-Ratio analysis of the aligned separations from (C, top) and (D, bottom).

shifting in the chromatographic separation, hence application of the TIC SF across all *m*/*z*. Additionally, one could imagine selecting a set of *m*/*z* ions to routinely check (as in Figs. 5A and 6A) how accurately the individual *m*/*z* SF compare to the TIC SF. The individual *m*/*z* should be selected to test key compound classes in the samples. If there is good accuracy between the individual *m*/*z* SFs relative to the TIC SF, then the TIC SF could more confidently be applied to all *m*/*z*, hence saving computational time.

4.3. Demonstration of TIF SF utilization by PCA classification

Retention time precision gained by using the TIC SF with sub-alignment, was evaluated using PCA. The scores plots from the PCA of the three gasoline sample types run over 2 days using the GC-TIC data before alignment is shown in Fig. 7A. Please note, we are initially evaluating the TIC SF alignment on the TIC data to serve as a benchmark to see what benefits there are to having the full GC–MS data aligned via the TIF SF approach. The only separation or "clustering" in the scores plot for the GC-TIC data is between the first two principal components (PCs) based on the temperature program type separation even though 95% of the variance is captured by the first two PCs. The first PC shows a separation between the P1 and P2 classes. A scores plot of the first two PCs of the same gasoline types after alignment of the GC-TIC data using the TIC SF are shown in Fig. 7B. The three types of gasoline samples all successfully cluster separately by gasoline type with no clear program type classification. This decrease in program type specification is important for proper classification.

The PCA of the full GC–MS data prior to alignment was very similar to Fig. 7A so is not shown for brevity. The PCA results of the full GC–MS data after TIC SF alignment are shown in a scores plot with PCs one and two in Fig. 7C. The type S samples clearly separate away from the types C and M samples on the first PC. However, the type C samples and type M samples are only slightly separated on the first two PCs even though the temperature program information is substantially removed. Fig. 7D shows more clearly how

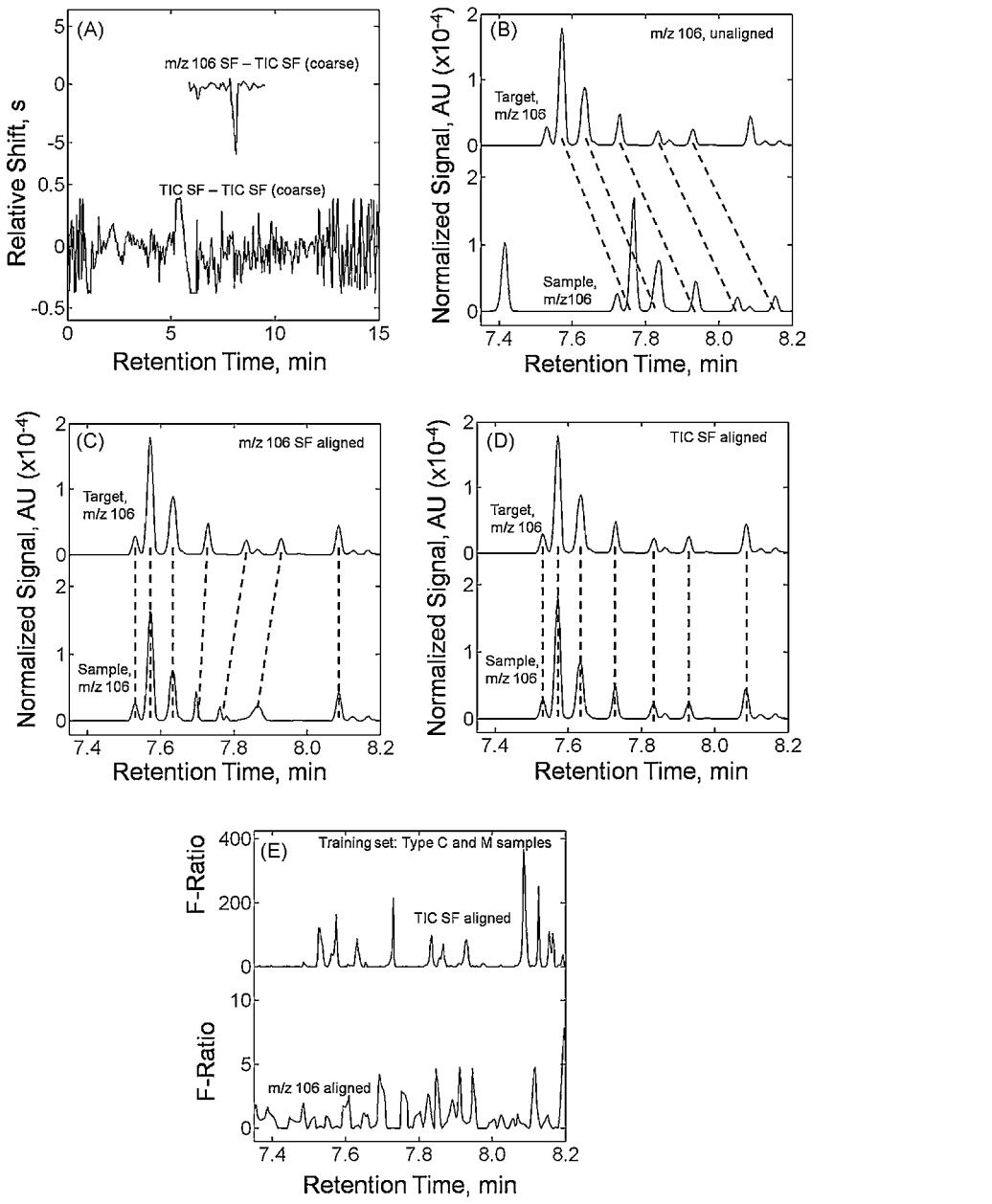

Fig. 6. (A) In contrast to Fig. 5A, the SF (with sub-alignment) for m/z ion 106, and the TIC SF (with sub-alignment) both relative to the TIC SF (without sub-alignment) show significantly different behavior over the entire signal range. (B) Data portion illustrating misalignment of the m/z 106 ion. (C) Misalignment as in (B) is not corrected by applying the m/z 106 SF with sub-alignment. (D) Misalignment as in (B) is corrected by applying the TIC SF with sub-alignment. (E) *F*-Ratio analysis of the aligned separations from (C, top) and (D, bottom).

well separated the sample types C and M are on the first two PCs. The degree-of-class separation (DCS) between these types is only 1.1, likely because of all the extra extraneous spectral information, whereas the DCS was 7.9 between the type C and M samples in Fig. 7B using the GC-TIC data. In both cases the DCS increased from an initial unaligned value of essentially zero (classified by program used instead). *F*-Ratios are a good way to find the regions of interest in a complex separation, especially if there is a second dimension of information as is the case with GC–MS, to fine tune the PCA data analysis to achieve a DCS more in favor of the analyst.

To evaluate the PCA results of both the GC-TIC and full GC–MS data, in the context of taking advantage of the information provided by the *F*-Ratio analysis using the type C and M training set, the ten best regions from the 2D *F*-Ratio analysis and the TIC *F*-Ratio analysis were used for comparison. The top ten *F*-Ratios are circled in Fig. 8A on the full 2D *F*-Ratio plot of the GC–MS data of sample types C and M. Fig. 8B shows the summation over all the mass channels from the 2D *F*-Ratios (from Fig. 8A) with the 1D *F*-Ratios using the GC-TIC data from sample types C and M. The 2D *F*-Ratio in Fig. 8A plot not only gives the time ranges of all the maximum *F*-Ratios, but the best m/z to use as well. The 2D information is lost when the summed 2D *F*-Ratios or the 1D *F*-Ratios are used to find regions of interest for enhancing classification studies. The identified m/z and times of interest were used to improve the PCA results of the full GC–MS data. The PCA scores plot in Fig. 8C displays the first two PCs and all three gasoline samples with 95% of the variance explained by these two PCs. The separation of types C and M is much greater, now with a DCS of 13.7, than in Fig. 7C and even better than the separation using the aligned TIC data shown in Fig. 7B. For comparison, the 1D TIC *F*-Ratio regions of interest identified in Fig. 8B were also used to evaluate the separation between sample types C and M using PCA, with the scores plot shown in Fig. 8D. However,

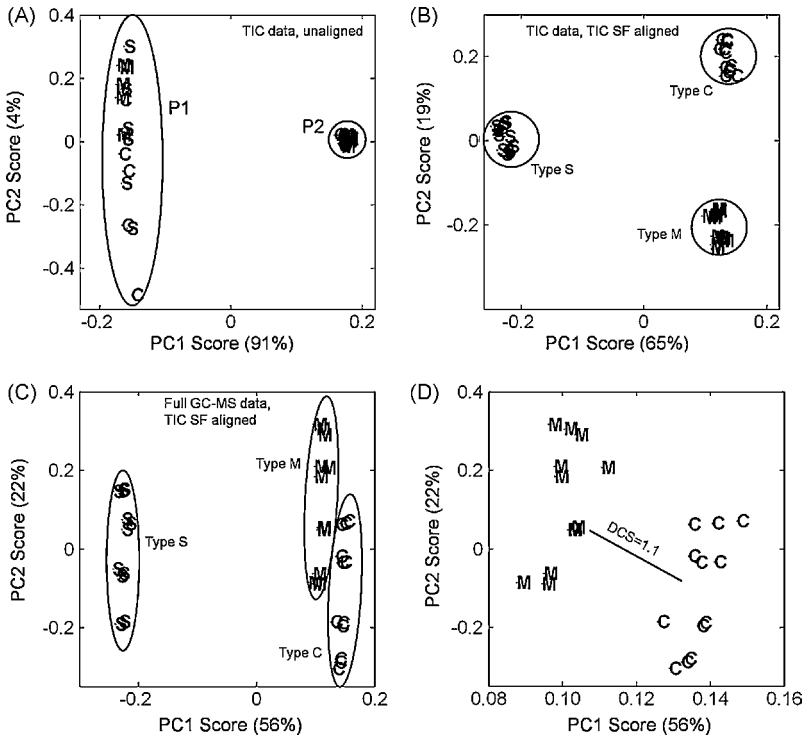

Fig. 7. Demonstration of PCA of the TIC chromatograms (as a benchmark of first order GC-TIC data) as compared to PCA of the entire GC–MS second order data that had each *m*/*z* aligned using the TIC SF methodology. (A) Scores plot from PCA with three gasoline types (C, M, and S) run on two temperature programs (P1 and P2) of GC-TIC data prior to alignment. Class separation is by separation program conditions. (B) Scores plot of GC-TIC data from (A) following complete alignment (with coarse and sub-alignment) shows proper fuel classification. (C) Scores plot, as in (A), but of the entire GC–MS data set prior to alignment. (D) Scores plot of entire GC–MS data set using TIC SF alignment with coarse and sub-alignment (e.g., Fig. 2C) shows proper fuel type classification. Note that all mass spectral information is preserved in this approach. A zoom-in of the two classes is provided.

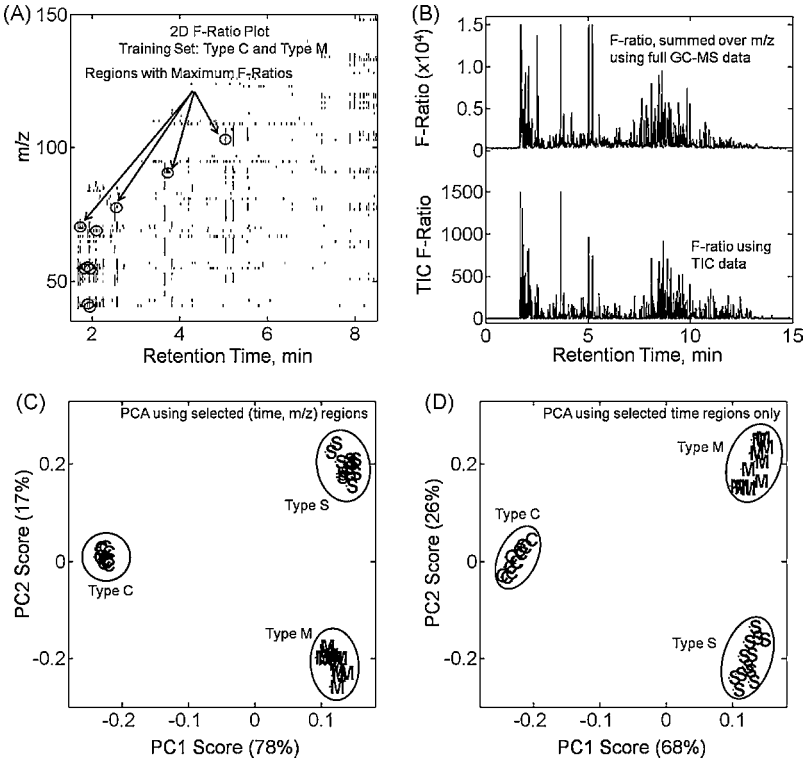

Fig. 8. Fisher Ratio analysis is used to improve the PCA classification by training from type C and type M fuels. (A) Fisher Ratio analysis of the full GC–MS data, as in Fig. 1B. (B) The *F*-Ratio analysis of the TIC data as in Fig. 1A (top) and the 1D projection of (A). (C) PCA scores plot using the top ten *F*-Ratios identified from (A) with PCA capturing 95% of the variance. (D) PCA scores plot using the top ten *F*-Ratios identified from (B, top) with PCA capturing 94% of the variance.

after the addition of the S type samples, the separation between all the classes remained essentially unchanged, with the DCS of 7.8 between the type C and M samples. The results show that the use of the TIC SF alignment procedure to the full GC–MS data (coarse and sub-alignment) to initially very misaligned data is essential in the subsequent classification using PCA, and the use of the *F*-Ratio approach further improves the classification.

5. Conclusions

The retention time precision of second order chromatographic data such as from GC–MS can be improved by using a TIC SF approach. An important point is that the column's condition must be relatively constant over the analysis (i.e., sufficiently constant run-to-run chemical selectivity). Significant changes in the chemical selectivity can cause more dynamic shifting between the individual m/z ions than can be sufficiently accounted for by the TIC SF. The TIC SF approach saves time over attempting to align all m/z individually or as a larger 2D windows of data, and indeed, aligning all m/z individually is not likely to work at many m/z that do not exhibiting sufficient signal (i.e., many peaks). The retention time alignment method described herein can be used for severely shifted data or data that are only slightly misaligned. For the larger shifts, larger windows are typically necessary to correctly align the chromatographic data. Because of the large shifts, a secondary, fine tuned alignment is necessary. Also, corrections to retain the area of peaks are necessary for alignment of severely shifted data. Because the chromatographic data is analyzed by a windowed approach and does not compare the profile of the data to every region in the target, the algorithm does not require a large amount of memory, and is very fast to implement.

Acknowledgements

The authors thank Karisa M. Pierce for developing the initial alignment algorithm, which was further developed and adapted in this report. This work was supported by the Internal Revenue Service through an Interagency Agreement with the U.S. Department of Energy. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO 1830. The views, opinions, or findings contained in this report are those of the authors and should not be construed as the official Internal Revenue Service position, policy, or decision unless designated by other documentation.

References

[1] D. Saison, D.P. De Schutter, F. Delvaux, F.R. Delvaux, J. Chromatogr. A 1216 (2009) 5061.
[2] R.M. Black, R.J. Clarke, D.B. Cooper, R.W. Read, D. Utley, J. Chromatogr. 637 (1993) 71.
[3] S. Prasad, K.M. Pierce, H. Schmidt, J.V. Rao, R. Guth, S. Bader, R.E. Synovec, G.B. Smith, G.A. Eiceman, Analyst 132 (2007) 1031.
[4] J.S. Ribeiro, F. Augusto, T.J.G. Salva, R.A. Thomaziello, M.M.C. Ferreira, Anal. Chim. Acta 634 (2009) 172.
[5] B.J. Prazen, K.J. Johnson, A. Weber, R.E. Synovec, Anal. Chem. 73 (2001) 5677.
[6] B.J. Prazen, C.A. Bruckner, R.E. Synovec, B.R. Kowalski, Anal. Chem. 71 (1999) 1093.
[7] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen, J. Chromatogr. A 1062 (2005) 113.
[8] A.M. Hupp, L.J. Marshall, D.I. Campbell, R.W. Smith, V.L. McGuffin, Anal. Chim. Acta 606 (2008) 159.
[9] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101.
[10] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111.
[11] D. Ballabio, T. Skov, R. Leardi, R. Bro, J. Chemometr. 22 (2008) 457.
[12] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.
[13] C. Christin, A.K. Smilde, H.C.J. Hoefsloot, F. Suits, R. Bischoff, P.L. Horvatovich, Anal. Chem. 80 (2008) 7012.
[14] S.J. Dixon, Y. Xu, R.G. Brereton, H.A. Soini, M.V. Novotny, E. Oberzaucher, K. Grammer, D.J. Penn, Chemom. Intell. Lab. Syst. 87 (2007) 161.
[15] C.G. Fraga, J. Chromatogr. A 1019 (2003) 31.
[16] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Anal. Chem. 73 (2001) 675.
[17] N. Pasadakis, E. Gidarakos, G. Kanellopoulou, N. Spanoudakis, Environ. Forensics 9 (2008) 33.
[18] A.E. Sinha, C.G. Fraga, B.J. Prazen, R.E. Synovec, J. Chromatogr. A 1027 (2004) 269.
[19] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1096 (2005) 177.
[20] J.H. Christensen, G. Tomasi, A.B. Hansen, Environ. Sci. Technol. 39 (2005) 255.
[21] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[22] K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1141 (2007) 106.
[23] R.G. Sadygov, F.M. Maroto, A.F.R. Huhmer, Anal. Chem. 78 (2006) 8207.
[24] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemometr. 20 (2006) 484.
[25] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231.
[26] L.F. Zhu, R.G. Brereton, D.R. Thompson, P.L. Hopkins, R.E.A. Escott, Anal. Chim. Acta 584 (2007) 370.
[27] A. de Juan, S.C. Rutan, R. Tauler, D.L. Massart, Chemom. Intell. Lab. Syst. 40 (1998) 19.
[28] S.J. Dixon, R.G. Brereton, H.A. Soini, M.V. Novotny, D.J. Penn, J. Chemometr. 20 (2006) 325.
[29] K.J. Johnson, B.J. Prazen, D.C. Young, R.E. Synovec, J. Sep. Sci. 27 (2004) 410.
[30] B.J. Prazen, C.A. Bruckner, R.E. Synovec, B.R. Kowalski, J. Microcolumn Sep. 11 (1999) 97.
[31] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, Anal. Chem. 77 (2005) 7735.
[32] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.
[33] J.T. Prince, E.M. Marcotte, Anal. Chem. 78 (2006) 6140.
[34] M. Kirchner, B. Saussen, H. Steen, J.A.J. Steen, F.A. Hamprecht, J. Stat. Software 18 (2007) 12.
[35] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.
[36] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 78 (2006) 2700.
[37] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.
[38] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833.
[39] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.
[40] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.
[41] W.G. Pool, J.W. deLeeuw, B. vandeGraaf, J. Mass Spectrom. 31 (1996) 213.