Journal of Chromatography A, 1218 (2011) 824–832

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Improved peak selection strategy for automatically determining minute compositional changes in fuels by gas chromatography–mass spectrometry

Jeffrey A. Cramer*, Nathan J. Begue[1], Robert E. Morris

U.S. Naval Research Laboratory, Chemical Sensing and Fuel Technology Section, Code 6181, 4555 Overlook Avenue, SW, Washington, DC 20375, USA

ARTICLE INFO

Article history:
Received 14 September 2010
Received in revised form 2 December 2010
Accepted 11 December 2010
Available online 21 December 2010

Keywords:
Gas chromatography–mass spectrometry (GC–MS, GCMS)
Parallel factor analysis (PARAFAC)
Analysis of variance (ANOVA)
Peak selection
Microbiological contamination (MBC)
Diesel fuel

ABSTRACT

During the development of automated computational methods to detect minute compositional changes in fuels, it became apparent that peak selection through the spectral deconvolution of gas chromatography–mass spectrometry (GC–MS) data is limited by the complexity and noise levels inherent in the data. Specifically, current techniques are not capable of detecting minute, chemically relevant compositional differences with sufficient sensitivity. Therefore, an alternative peak selection strategy was developed based on spectral interpretation through interval-oriented parallel factor analysis (PARAFAC). It will be shown that this strategy outperforms the deconvolution-based peak selection strategy as well as two control strategies. Successful application of the PARAFAC-based method to detect minute chemical changes produced during microbiological growth in four different inoculated diesel fuels will be discussed.

Published by Elsevier B.V.

1. Introduction

In many areas of fuel study, it is often necessary to examine multiple fuel samples with the goal of accurately determining their compositional similarity or dissimilarity. Such a capability would directly facilitate research in materials compatibility, fuel thermal and storage stability, and responses to environmental exposure. As an example of the latter topic, the U.S. Navy frequently stores its shipboard mobility fuels in seawater-compensated tanks, allowing marine vessels to maintain ballast and stability as fuels are consumed. This type of storage provides a significant opportunity for microbiological contamination (MBC) from microorganisms introduced from the seawater to metabolize fuels and render them unusable or otherwise compromised. As there is a great deal of interest in exploring non-petrochemical, i.e. alternative, fuel usage on board Naval ships, we have been engaged in a study to determine how these fuels are metabolized as a consequence of MBC in seawater-compensated tanks. To this end, a sensitive analytical strategy was needed to elucidate the relatively small changes that occur at the fuel-water interface.

Such an analytical strategy must not only be comprehensive and accurate but must also be as automated as possible to facilitate regular fuel analysis. Such a method consequently requires a high level of analytic robustness, i.e. a built-in resistance to false positive and false negative results, to allow for its practical and confident application by non-expert users. Fortunately, a great deal of information regarding fuel composition and performance can already be obtained from gas chromatography–mass spectrometry (GC–MS) [1–3] data, making it an ideal analytical technique upon which to base an overall analysis strategy. Furthermore, a comprehensive automated toolkit for interpreting GC–MS data is already available in AMDIS (the Automated Mass Spectral Deconvolution and Identification System) [4], provided by the National Institute of Standards and Technology (NIST) [5]. AMDIS functions by selecting chromatographic peaks from the total ion chromatogram (TIC) obtained from the summed GC–MS data. This peak selection follows from a spectral deconvolution that is, in itself, reliant upon a noise analysis and subsequent noise-compensated component perception. Spectral peaks that are identified through deconvolution are then used to choose the most relevant mass spectral profiles and, in turn, identify the components represented by them. It should be noted that fuel-derived GC–MS data have previously been analyzed successfully with the AMDIS toolkit [6], rendering its use in the present context, at least theoretically, a routine application. Unfortunately, AMDIS deconvolution-based algorithms met with limited success when applied to the noisy, complex data that were collected.

To address the limitations of deconvolution-based peak detection in such unfavorable GC–MS data, an alternative strategy was developed, based on a well-established technique from the field of

* Corresponding author. Tel.: +1 202 404 3419.
E-mail address: jeffrey.cramer@nrl.navy.mil (J.A. Cramer).

[1] [1]National Research Council (NRC) Post-Doctoral Fellow.

0021-9673/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.chroma.2010.12.037

chemometrics known as parallel factor analysis (PARAFAC) [7]. The use of PARAFAC was inspired by its ability to extract unique data profiles from three- or higher-dimensional data sets [8] that can be subsequently used to produce practical, quantitative predictions [9–11]. Extracted data profiles, based only on the chemically relevant covariances common to all of the GC–MS spectra, effectively minimize the effects of competing analyte diversities and instrument noise in a comprehensive manner. In this instance, PARAFAC functions in much the same way as the deconvolution-based peak selection methodology, except that selected peak locations are based directly upon underlying linear data variance, as opposed to derived data structures. It should be noted that two conceptually similar techniques have previously appeared in the literature, the first applying the Generalized Rank Annihilation Method (GRAM) [12] to gas chromatography–mass spectrometry data as produced using selected-ion monitoring (GC-SIM) [13], and the second combining GRAM with PARAFAC to form the hybrid technique of GRAM-PARAFAC [14]. The use of GRAM in these two analysis strategies enhances the resolution of individual components, but the fundamental algorithm, as is also the case with the previously published DotMap algorithm [15], requires the collection of pure component spectra to achieve the enhancement. In the present circumstances, the sheer number of compounds that can potentially be present in fuels renders the collection of pure component spectra unrealistic.

Regardless of PARAFAC's utility, however, the large amount of numerical data present in the available GC–MS chromatograms will tend to exceed the practical limits on the available time and computer resources required to perform even a basic PARAFAC analysis. In addition, the number of spectral variances that are not co-linear across all retention times and mass/charge (m/z) ratios would hinder the derivation of underlying linear variances across the entire data set simultaneously. To address both of these challenges, an approach was developed that subdivides the parent data set and executes a series of PARAFAC computations on smaller sequential portions of the original GC–MS data cube. Mass spectra, in this case, are selected through the use of "windows," each consisting of a predefined number of retention times, that can be visualized as "moving" across the retention time axis. PARAFAC is performed repeatedly upon the smaller three-dimensional data cubes defined within these windows, and models are built to represent the most significant underlying co-linear information in both the local retention times and MS spectra as well as the distribution of relevant information amongst the samples. The maxima of the PARAFAC retention time results are first used to select the most appropriate retention times, as an alternative to the deconvolution-based peak selection. The mass spectral PARAFAC results corresponding to selected retention times are then used to select compounds exactly as if they were true mass spectra. Finally, normalized peak area results are used to scale derived amounts of compounds appropriately for each sample to determine relative increases and decreases in individual compounds between the samples represented within the original data cube.

It is interesting to note that this technique's development has independently resulted in a final form similar to another PARAFAC-based analysis strategy that has previously appeared in the literature [16,17]. The differences between the previously published and presently proposed strategies are found in how they accumulate and validate results. Specifically, the previously published strategy focuses upon the use of multiple differently sized, and potentially quite large, PARAFAC models and the use of diagnostic values including match factor to internally compare the mass spectral results obtained from multiple models to assure the quality of results. The present technique, by contrast, produces consistently small PARAFAC models that vary in terms of their location instead of their complexity. These models represent less spectral variance individually, but the use of many such overlapping models using different retention time scales both captures all useful data variance and internally validates said variance without the use of larger models or match factor values.

The effectiveness of the presently proposed interval-oriented PARAFAC strategy to determine the impact of MBC on diesel fuel composition is evaluated in the present work in terms of competing and control strategies as well as the effects of modifying experimental parameters. Then, the strategy will be applied to a variety of diesel fuels to ascertain the chemical changes that occurred as a consequence of MBC.

2. Materials and methods

2.1. Fuel samples

Initial GC–MS data, reported upon previously [18], were collected from four different diesel fuels: a specification F-76 petrochemical diesel fuel; an ultra-low sulfur diesel (ULSD) petrochemical fuel; the same ULSD fuel containing 5% fatty acid methyl ester (FAME) biodiesel (B5); and a Fischer–Tropsch (FT) synthetic diesel fuel. For each fuel, five replicates was collected for each of two conditions: (1) an abiotic control set consisting of synthetic seawater spiked with a relatively low dose of poisoned microorganisms with microorganism-promoting nutrients, and (2) an experimental set consisting of synthetic seawater containing a relatively high dose of microorganisms and microorganism-promoting nutrients. After 25 days of exposure, the nutrients, microorganisms, and poisons were removed from both sample sets prior to GC–MS analysis to ensure that only the consequences of MBC on fuel composition were assessed.

2.2. GC–MS analysis

Data were collected on an Agilent 5890 GC with an Agilent 5971 mass selective detector. Samples were diluted to 1:100 in dichloromethane, and injections of 1.0 μL were made with an autoinjector. An AT-1 capillary column (50 m × 0.25 mm ID, 0.20 μm film thickness) was used with an oven temperature program that initiated data collection at a temperature of 40 °C and ramped at 10 °C/min to 290 °C, holding this temperature for the remaining duration of the data collection. Data were collected at column retention times from 6.8 to 36.1 min at a frequency of 2.8 mass spectra per second, over 40–279 m/z. It should be noted that this data tended to show peak widths of between 7 and 9 variables, or about 2–3 s, along the retention time axis.

3. Calculation

Analysis strategies are assessed in the present work by their ability to determine which compounds within the fuel samples increased and decreased as a consequence of MBC. Therefore, in the following text, if compounds are detected at a greater abundance in the control samples, they are said to be decreasing, and compounds that are more abundant in the MBC experimental samples are said to be increasing.

All GC–MS data were imported into MATLAB R2010a (MathWorks, Inc., Natick, MA) and assembled into four data cubes, each representing one of the different diesel fuels. Data deconvolution, PARAFAC, and other peak-picking algorithms, as well as preprocessing algorithms, were developed and performed within the MATLAB environment with functionality provided by the PLS_Toolbox for MATLAB ver. 4.2 (Eigenvector Research, Inc., Wenatchee, WA). The PARAFAC algorithm was performed under the constraints of non-negativity and orthogonality [8] unless otherwise stated. It should be noted here that the alternative approaches to peak selection, all fundamentally based on modifications to the overall AMDIS methodology, still make use of the same compound identification algorithms developed within the context of the AMDIS algorithm toolkit. These identification algorithms were used in concert with the alternative selection strategies in the same manner as they would have been used in a typical AMDIS-based analysis. Specifically, mass spectral data indicated by the peak selections from all techniques are identified using the NIST Mass Spectral Search Program for the NIST/EPA/NIH Mass Spectral Library, version 2.0f, build October 8, 2008 [19].

Before analysis, each raw sample chromatogram was first normalized to unit area. Since the peaks relevant to each fuel did not extend past 25 min retention time, the inclusion of the longer tail in this step helped to mitigate normalization-based data distortions. After normalization, each ten-sample fuel data cube underwent a self-contained correlation optimized warping (COW) procedure [20] with a window size of 100 data point variables and a slack length of 10 data point variables. It should be noted here that, based on the frequency given previously, the total GC axis window size corresponds to about 36 s worth of mass spectral data. The COW preprocessing ensured that covariant chemical components would be detected as such by correcting for minor shift variations across the retention time axis.

The deconvolution-based peak selection procedure used in AMDIS was reproduced in the MATLAB programming environment to allow for fine control over experimental parameters and variables. The AMDIS toolkit maintains several parameters that can and should be selected and, in many cases, adjusted by end users to optimize results, as indicated in the published work [4] upon which AMDIS is based. For instance, the mass spectrometer abundance threshold, i.e. the signal intensity that must be exceeded for a mass spectral value to be saved to the instrument, was set to a constant 150 *m*/*z*, a value established during instrument tuning. The present work also evaluates adjustments in the maximum number of variables on each side of a potential GC peak to use for deconvolution, which was set to 12 in the aforementioned published work [4]. A rejection threshold was additionally implemented to determine if possible peaks are relevant based on their height, which the previously published work defaults to 4× the noise level and the present work adjusts to optimize results. Finally, the aforementioned match factor (MF) value threshold produced by the NIST Mass Spectral Search Program can be used to quantify the likelihood that the mass spectrum was identified correctly and reject results if necessary. Although MF was defined as a value between 0 and 100 in the original reference, the Mass Spectral Search Program reports the value as a proportional value ten times greater, i.e. 0–1000. For the purposes of direct comparison, these larger values are scaled back to the original 0–100 range in the present work. It should be noted here that, although MF values are always incidentally calculated when using the Mass Spectral Search Program, they are actually used to quantify goodness of fit only when evaluating the deconvolution-based peak selection methodology.

In performing the deconvolution-based peak selection algorithm, the individual GC–MS replicate chromatograms were necessarily assessed separately. An overall set of combined results for each ten-sample fuel population was obtained through a two-step procedure. First, whenever a component was detected multiple times within a single GC–MS chromatogram, the results of these multiple detections were added together to form a single, unique component result for each sample. It should be noted here that this step, utilizing results obtained directly from the Mass Spectral Search Program, only added together those components with identical names, and components that are differently named isomers of each other were not added together. Second, these unique component results, across all samples, were sorted from their highest to lowest values. This sorting occurred separately for the control and MBC sample sub-populations. From these overall results, changes in component content were identified if the compared maxima and minima of the control and MBC results could reliably accommodate the identification. Thus, a decrease between the control and experimental samples was reported for a given compound if the minimum component content amongst the five control samples was greater than the maximum component content amongst the five MBC samples, and vice versa. The most accurate compound assessments obtained by using this algorithm with multiple parameter settings are appropriately comparable to the PARAFAC-based results and associated control trials.

PARAFAC was performed repeatedly upon the three-dimensional data cubes defined by each ten-sample data set as they are truncated along the retention time axis to the size of a "window" at a series of positions. As a consequence, window size and window positioning (or "movement" as defined by a regular step size) are two scalable parameters that must be assessed during the course of this work. To this end, the PARAFAC-based peak selection is performed under two separate sets of parameters, denoted in the remainder of the document as a 'thorough' version of the technique (window sizes of 5, 100, and 300 data points, step size 1 data point in all cases) and a 'fast' version of the technique (window sizes of 25 and 150 data points, step sizes 20 and 10 data points respectively) based on their respective advantages. It should be noted here that these window sizes correspond to about 2, 36, and 107 s worth of GC–MS data in the case of the thorough technique, and 9 and 53 s worth of GC–MS data in the case of the fast technique. The GC PARAFAC results collected using each window size/step size combination within each version of the technique are compiled to produce a single list of selected peaks. The mass spectral PARAFAC results corresponding to these peaks are then assessed using the Mass Spectral Search Program, and the relative amounts of each compound for each sample are also obtained by calculating the area defined within each window. These relative compound contents are then added together into an overall result set and used to determine increasing and decreasing compound contents, as percentages, between the two sample populations as described for the deconvolution-based analysis. It should be noted that percent changes are always calculated as a change from the lower component content to the higher component content, regardless of whether or not the components are higher before or after MBC, to maintain the same relative scale between components that are increasing and decreasing.

Two additional control strategies were also performed to further evaluate the performance of the PARAFAC-based peak selection. In the first control strategy, PARAFAC results were abandoned in favor of the maximum total ion chromatograph (TIC) value to be found within each window size/step size combination. As this does not yield a complementary set of PARAFAC modeling results, average mass spectra were first used to derive the component information common to all of the spectra. Relative quantities of each component were then derived, as with the interval PARAFAC technique, by using the area of each normalized GC–MS spectrum of an individual fuel sample within the given window. In the second control strategy, peak selection was abandoned entirely, and all available retention times were evaluated in terms of MS spectral content and GC–MS spectral area. This last control strategy, which utilized no initial input from the GC axis, necessarily assumes that compound identifications based only on noise and other spectral artifacts will not be consistent across all available MS spectra.

It was quickly determined that one-factor PARAFAC (or 1-PARAFAC) models were the most useful within the context of the present work. Not only were one-factor models slightly faster to derive than larger models, but larger models were also not guaranteed to produce repeatable results, as would be required from

Table 1
Automated GC–MS B5 fuel composition results using deconvolution-based peak selection. Column labels identify analysis parameters as maximum variable range/noise-based rejection threshold. Results obtained with a match factor threshold of 75 are reported in parentheses only if they differ from results obtained using a value of 0.

	6/4×	12/4×	18/4×	24/4×	30/4×	36/4×	50/4×
All Confirmed							
Identified Negatives	0	0	0	0	0	0	0
False Positives	0	0	0	0	0	0	0
All FAME							
Identified Negatives	1	1	0	0	0	0	1
False Positives	0	0	0	0	0	0	0
	6/8×	12/8×	18/8×	24/8×	30/8×	36/8×	50/8×
All Confirmed							
Identified Negatives	1	1	1	1	1	1	0 (1)
False Positives	0	0	0	0	0	0	0
All FAME							
Identified Negatives	1	2	2	2	2	2	1 (2)
False Positives	0	0	0	0	0	0	0
	6/20×	12/20×	18/20×	24/20×	30/20×	36/20×	50/20×
All Confirmed							
Identified Negatives	1	1	1	1	1	1	1
False Positives	0	0	0	0	0	0	0
All FAME							
Identified Negatives	1	1	2	2	2	2	2
False Positives	0	0	0	0	0	0	0
	6/50×	12/50×	18/50×	24/50×	30/50×	36/50×	50/50×
All Confirmed							
Identified Negatives	1	1	1	1	1 (0)	1 (0)	1 (0)
False Positives	0	0	0	0	0	0	0
All FAME							
Identified Negatives	1	1	2	1	1 (0)	1 (0)	1 (0)
False Positives	0	0	0	0	0	0	0
	6/100×	12/100×	18/100×	24/100×	30/100×	36/100×	50/100×
All Confirmed							
Identified Negatives	0	0	0	0	0	0	0
False Positives	0	0	0	0	0	0	0
All FAME							
Identified Negatives	0	0	0	0	0	0	0
False Positives	0	0	0	0	0	0	0

an automated analysis strategy. This lack of repeatability is a consequence of PARAFAC's Alternating Least Squares (ALS) algorithm when it is initiated with random variables and carried out under the aforementioned non-negativity and orthogonality constraints. This ALS algorithm is used to perform PARAFAC in the absence of *a priori* knowledge [8], which is necessary to maintain the technique's general applicability. The ALS algorithm arrives at a conclusion as to how much linear data variance can be explained with each linear factor, and this conclusion arrives at this conclusion differently depending upon these initial random variables. Although a trained operator can decide upon the accuracy of a particular result, this would not apply if this procedure was to be automated for non-expert use. In comparison, repeatability and reliability are assured with a 1-PARAFAC modeling approach because the ALS algorithm can only reach one unambiguous conclusion in the absence of multiple potential linear factors. Although it initially seems counterintuitive to use such a small data model for such a large data set, the sheer number of 1-PARAFAC models that are created during the course of the interval-oriented strategy more than compensates for individual model inadequacies. It should also be noted that the production of 1-factor PARAFAC model from normalized, non-negative data renders the aforementioned constraints of non-negativity and orthogonality superfluous, as the lone factors can neither be negative nor non-orthogonal to other factors. For the sake of a comprehensive evaluation, the results obtained when using larger PARAFAC models will also be reported in the form of three replicates per factor increase, and additional results will be obtained without the orthogonality constraint in place.

Evaluations of the analysis strategies in the present work were initially accomplished by focusing upon the analysis results obtained from the B5 fuel since a list of components known to disappear in 20% biodiesel fuel blends upon MBC is available in the literature [21]. The B5 data set was analyzed using all versions of the four techniques, and results were collected to determine how many of the compounds that are known to be consumed (All Confirmed) were, in fact, found to be at lower concentrations compared to the control samples (Identified Negatives) as well as how many were falsely indicated to increase (False Positives). In addition, Identified Negative and False Positive results were also obtained assuming that each FAME constituent determined to be decreasing during the course of the analysis was properly detected as such, and that all FAME constituents determined to be increasing were erroneous. This was based on the fact that the metabolic breakdown of every individual FAME constituent is very similar [22], and that the metabolic consumption of the FAME compounds listed in the literature strongly implies a decrease in other FAME compounds upon MBC.In the context of a conservative automated analysis, techniques are considered to be more successful when Identified Negative results are high and False Positive results are as low as possible. Once an analysis strategy was deemed to be acceptable in accordance with these criteria, the specification F-76, ULSD, B5, and FT diesel data sets were fully analyzed using that strategy.

Table 2
Results of PARAFAC-based peak selection algorithm and two control algorithms on B5 fuel composition results.

	All mass spectra	Local TIC maximum (thorough)	Local TIC maximum (fast)	Local 1-PARAFAC (thorough)	Local 1-PARAFAC (fast)
All Confirmed					
Identified Negatives	5	8	7	6	6
False Positives	4	9	6	3	1
All FAME					
Identified Negatives	11	9	5	15	9
False Positives	1	1	1	1	0
Running Time (s)	5243	15,870	732	24,360	978

4. Results and discussion

4.1. Parameter selection in deconvolution-based peak selection

Results of the deconvolution-based GC–MS evaluations, using several different combinations of maximum variable ranges and noise-based rejection threshold levels, can be found in Table 1. The MF threshold was also used for each listed set of results, at both 0 and 75, though, as reported in the table, the altering of this threshold added very little discrimination to the overall analysis. As can be clearly seen, the number of Identified Negatives, for either Confirmed or All FAME compounds, is quite low regardless of the number of variables used to deconvolute the peaks or the noise threshold. The best results to be found in the table, i.e., simultaneous detection of one Confirmed compound and two FAME compounds, indicate in every case the detection of two methyl esters, with one of these results, methyl hexadecanoate or hexadecanoic acid methyl ester, confirmed in the literature and additionally reported as a Confirmed result. It should be noted that the remaining results in the table indicate either methyl hexadecanoate or the other methyl ester detected in the "best" results, linolenic acid or (z,z,z)-9,12,15-octadecatrienoic acid methyl ester. The results in Table 1 also indirectly illustrate the effects of high levels of data complexity and noise on a deconvolution-based peak selection strategy, as the 4× rejection threshold, found to be adequate in the original reference, performs poorly in the present circumstances.

The Identified Negative and False Positive results were partially reproduced using the standard AMDIS software package at its default settings, which include MF corrections yielding a "net" MF value. This software makes use of the original, published scale for MF results, i.e. 0–100, with larger numbers denoting greater probabilities of a match. Using the software, methyl hexadecanoate was the only compound reliably identified in all of the control GC–MS spectra with net MF values of 86, 84, 83, 85, and 85, and two of the five experimental spectra with net MF values of 81 and 82. No compounds were reliably identified in the remaining three experimental spectra, with net MF values of 66, 70, and 68 for methyl hexadecanoate.

The AMDIS deconvolution-based algorithms used for peak selection, then, were found to be insufficiently accurate in detecting very small compositional differences in diesel fuels that had undergone MBC. These initial results indicated that the complexity and noise levels inherently present in this GC–MS data hindered the ability of the AMDIS toolkit to accurately select individual TIC peaks. The major limitation of deconvolution in this context is scale, as fuels produce a complex and noisy background against which the relatively minor peak changes between control and MBC-affected samples are quite difficult to detect.

4.2. Analysis strategy selection

The All Confirmed and All FAME results obtained by using the alternative peak selection strategies, i.e. the two control strategies and the PARAFAC-based strategy, are shown in Table 2. With respect to confirming the loss of the compounds listed in the literature for MBC growth, the use of the local TIC maximum for mass spectra selection actually provided the most Identified Negative results, but at the expense of detecting an inordinate number of False Positives. False Positive results themselves are minimized through the use of the local 1-PARAFAC modeling strategy, and the Identified Negative results for this technique are superior to those obtained when using all mass spectra indiscriminately. The PARAFAC-based modeling strategy also found the largest number of FAME compounds decreasing in relative content between the control and experimental data sets, although it should be noted that a False Positive does appear in these results. Although the thorough version of the Local 1-PARAFAC peak selection strategy required the most computational time to reach completion, the fast version was completed in just a little over 16 min.

Based on these results, the PARAFAC-based peak selection strategy was the most promising of the alternative peak selection strategies, especially if analysis time is not considered a critical factor. Furthermore, when comparing the results in Table 2 to those found in Table 1, it is clear that the best deconvolution-based peak selection strategy did not perform as well as the PARAFAC strategy.

Regardless of the promising results shown in Table 2, the reporting of even a few false positive results is deemed unacceptable for the purpose of developing an automated analysis tool that could be relied upon by non-expert users. Furthermore, false positive results in the present circumstances are seen as more detrimental to an effective analysis than false negative results because false positive results are obtained despite the known decreases occurring in the detected component amounts. Initially, the ideal solution to filter out erroneous PARAFAC-based results appeared to be the use of MF values that are already calculated by AMDIS. Unfortunately, low MF results do not consistently correlate with False Positive results, as can be seen in the All Confirmed results from the thorough version of the Local 1-PARAFAC in Fig. 1. Amongst the nine Confirmed results found through the PARAFAC-based peak selection strategy, MF results varied widely enough amongst the False Positive results so as to make them indistinguishable from the Identified Negative results. It should again be noted that MF values have been incorporated in result validation in the aforementioned similar PARAFAC-based techniques [16,17], but can be seen to be insufficient for the present work.

The figure also shows the absolute percent changes that are calculated between the control and experimental fuel populations for the nine Identified compounds. These percent changes are calculated from the same maxima and minima used to initially determine if there is a significant difference in the content of a particular compound between the control and MBC populations. This information is provided to show that the False Positive results indicate relatively small changes compared to most of those indicated by the Identified Negative results. This is reasonable, since the False Positive results are based on incorrectly interpreted non-chemical signals in the GC–MS data and would therefore be less likely to be perceived as the same magnitude as the Negative results that are

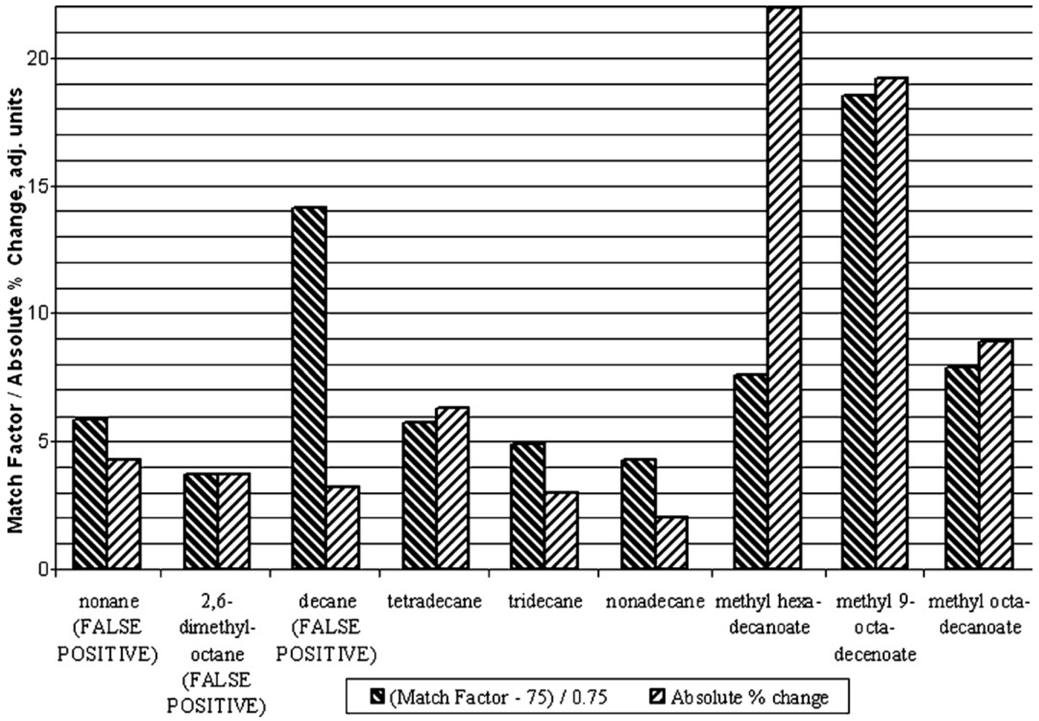

Fig. 1. Adjusted match factor and percent change results derived from the nine Confirmed compounds identified through the use of the thorough version of the 1-PARAFAC peak selection strategy.

based on actual underlying chemical changes. This initially indicates that there is merit in developing a validation parameter that accepts or rejects the presence of a given compound depending on the magnitude of any detected increase or decrease. However, in the example shown in the figure, setting such a parameter at 5%, which would accurately reject the three False Positive results, would also reject the correct identification of tridecane and nonadecane, and reducing the sensitivity of an analytical technique to provide robustness is, of course, undesirable. Therefore, it is seen here that readily available metrics for result validation are not well-suited to the present challenge, and a new metric can and should be introduced to better accommodate the results produced by the multiple 1-factor PARAFAC models.

In order to better preserve the correct Identified Negative results while still discriminating against False Positive results, an alternative threshold-based strategy was developed. Specifically, analysis of variance (ANOVA) [23] was used to determine if the results obtained from using either the PARAFAC algorithm or the control strategies' averaged mass spectra are, in fact, statistically distinct from random data variance. A one-way ANOVA test was used to determine if the control and MBC sample populations, as defined by PARAFAC or area results, were statistically distinct populations. This evaluation was performed repeatedly with respect to the compound identified at each window/step combination. To attain diagnostic information distinct from the area-based difference results and increase robustness, the technique was applied to the PARAFAC results corresponding to the sample axis. Compound results were only allowed to proceed further in the algorithms if it was determined that the differences found between the sample populations, as defined through PARAFAC, were statistically distinct from random noise to within a 99.99% confidence interval. The results of using this one-way ANOVA assessment as a filtering step in the alternative peak selection strategies are shown in Table 3.

The results in Table 3 confirm the effectiveness of the ANOVA filtering step to significantly decrease the number of False Positive results in almost all cases. This also provides a significant improvement in calculation times in almost all cases, likely a result of the decreased use of the Mass Spectral Search Program. Unfortunately, particularly in the case of the Local TIC Maximum control strategies, the number of Identified Negative results also decreased, indicating that their presence in the original results was in error. An interesting, albeit unintended, benefit is also seen in the increase in Identified Negative results obtained in a few cases. This increase is a consequence of the removal of at least some of the misleading intermediate results used to produce the final results. If, for example, an Identified compound is determined to be decreasing in one window/size combination but increasing in another, then the dis-

Table 3
Results of PARAFAC-based peak selection algorithm and two control algorithms, with an additional ANOVA result-filtering step, on B5 fuel composition results.

	All mass spectra	Local TIC maximum (thorough)	Local TIC maximum (fast)	Local 1-PARAFAC (thorough)	Local 1-PARAFAC (fast)
All Confirmed					
Identified Negatives	7	5	4	6	5
False Positives	0	1	0	0	0
All FAME					
Identified Negatives	9	8	4	16	9
False Positives	1	0	0	0	0
Running Time (s)	5078	14,853	731	20,023	847

Table 4
PARAFAC-based analysis results obtained by using multiple underlying factors and the orthogonality constraint when selecting peaks during B5 fuel composition determination.

	2 Factors (Rep. 1)	2 Factors (Rep. 2)	2 Factors (Rep. 3)	3 Factors (Rep. 1)	3 Factors (Rep. 2)	3 Factors (Rep. 3)
All Confirmed						
Identified Negatives	6	5	4	5	5	5
False Positives	2	0	1	0	1	1
All FAME						
Identified Negatives	4	4	6	4	4	4
False Positives	2	1	2	1	1	2
Computational Time (s):	865	848	855	902	917	901

tribution of the maxima and minima across both the control and MBC sample populations would be significantly affected. If, on the other hand, one of these two sets of intermediate results is filtered out using ANOVA, then a final trend becomes more apparent.

The use of ANOVA removes all False Positives from the Local 1-PARAFAC results, and almost all False Positives from the two control strategies. Although ANOVA actually allows the control strategy involving All Mass Spectra to be slightly more effective than the Local 1-PARAFAC strategies with respect to Confirmed results, the sole remaining FAME False Positive result and the sheer number of FAME compounds identified by the thorough 1-PARAFAC algorithm still demonstrate the greater utility of the PARAFAC-based peak selection. In the context of a conservative automated analysis, PARAFAC combined with ANOVA, especially the thorough version of the strategy, produces the most useful, reliable results.

4.3. *PARAFAC models using multiple factors*

The results in Tables 4 and 5 show the results of determining multiple simultaneous underlying linear factors through the PARAFAC peak selection algorithms, with and without the orthogonality constraint, respectively. These results were collected by mimicking the fast version of the original PARAFAC-based algorithm in most respects and simply increasing the number of factors to include in each model. Individual compound results from each factor were first collected into a single set as if they were derived from a single factor, and then processed normally. The lack of confident repeatability necessitates the collection of replicates for each number of factors to more completely represent the desired information.

Table 4 shows component results calculated with the orthogonality constraint in place. Although there is a slight increase in computation time when increasing the number of factors, the more striking trend is the lack of repeatability in obtaining both All Confirmed and All FAME results, either with respect to Identified Negatives or False Positives. Table 5, by contrast, shows the two-factor model results recalculated with the orthogonality constraint lifted. The lack of this constraint improves results considerably in terms of both repeatability and thoroughness, which is consistent with the in-literature observation that orthogonality is difficult to find in chromatographic data [8,24]. In essence, lifting the orthogonality restraint allows PARAFAC to have the flexibility to find more appropriate underlying data factors in GC–MS data. However, not only are the All FAME results still not as large as when 1-PARAFAC models are used, but the times required to find underlying factors with this newfound flexibility increase dramatically. These time increases, in turn, could render the thorough version of the technique impractical if time is a consideration, as the thorough version of the single-factor strategy can be performed in less time than the fast version of the multifactor strategy.

Table 5
PARAFAC-based analysis results obtained by using multiple underlying factors and no orthogonality constraint when selecting peaks during B5 fuel composition determination.

	2 Factors (Rep. 1)	2 Factors (Rep. 2)	2 Factors (Rep. 3)
All Confirmed			
Identified Negatives	8	7	8
False Positives	0	0	0
All FAME			
Identified Negatives	7	8	8
False Positives	0	0	0
Computational Time (s):	26,450	26,348	26,365

It may be possible to modify the PARAFAC-based peak selection algorithm in such a way that the lack of repeatability would be somewhat or perhaps even completely mitigated, such as with result averaging, more stringent convergence parameters, or the use of non-random initial values in the ALS algorithm. It may also be possible to increase algorithm speed to allow for the more practical removal of the orthogonality constraint to further improve results. However, the results in Tables 4 and 5 do not indicate that such modifications would provide significant improvements in the component results obtained using 1-PARAFAC combined with ANOVA as reported in Table 3.

Table 6
"Fast" ANOVA-augmented 1-PARAFAC analysis results obtained from the B5 data set.

B5 (5% biodiesel/95% ULSD)	Percent increase (+) or decrease (−)
1,1-cyclobutanedicarboxamide, 2-phenyl-n,n′-bis(1-phenylethyl)-	5.92
1,4-benzenediol, 2,6-bis(1,1-dimethylethyl)-	12.07
2-cyclohexen-1-ol, 2-methyl-5-(1-methylethenyl)-	4.78
benzene, 1-ethyl-3-methyl-	6.27
cyclotrisiloxane, hexamethyl-	72.02
1-heptatriacotanol	−2.27
1-hexadecanol, 2-methyl-	−5.55
2,5-octadecadiynoic acid, methyl ester	−0.52
2-butyloxycarbonyloxy-1,1,10-trimethyl-6,9-epidioxydecalin	−1.73
2-dodecen-1-yl(-)succinic anhydride	−1.42
2-piperidinone, n-[4-bromo-n-butyl]-	−1.98
7-heptadecene, 17-chloro-	−1.46
9,12,15-octadecatrienoic acid, . . . ethyl ester, (z,z,z)-	−0.32
9,12-octadecadienoic acid, methyl ester, (e,e)-	−37.98
9,12-octadecadienoyl chloride, (z,z)-	−35.92
9-octadecenoic acid (z)-, methyl ester	−18.89
cyclopropanebutanoic acid, . . ., methyl ester	−16.47
cyclopropanedodecanoic acid, 2-octyl-, methyl ester	−4.86
cyclopropanepentanoic acid, 2-undecyl-, methyl ester, trans-	−7.56
dodecane, 2,6,10-trimethyl-	−4.31
falcarinol	−0.47
hexadecane	−5.76
hexadecanoic acid, 14-methyl-, methyl ester	−6.48
hexadecanoic acid, methyl ester	−49.13
nonadecane	−4.85
octadecanoic acid, methyl ester	−28.30
tert-hexadecanethiol	−4.83
tetradecane, 2,6,10-trimethyl-	−4.63

Table 7
"Fast" ANOVA-augmented 1-PARAFAC analysis results obtained from the specification F-76 diesel data set.

Specification F-76 diesel fuel (petrochemical)	Percent increase (+) or decrease (−)
1,1-cyclobutanedicarboxamide, 2-phenyl-n,n′-bis(1-phenylethyl)-	20.80
10,13-octadecadiynoic acid, methyl ester	16.64
10-heptadecen-8-ynoic acid, methyl ester, (e)-	10.98
12,15-octadecadiynoic acid, methyl ester	4.01
1-decen-4-yne, 2-nitro-	18.24
1h-2,8a-methanocyclopenta[a]cyclopropa[e]cyclodecen-11-one, . . .	2.15
1-octadecanesulphonyl chloride	5.00
2,5-octadecadiynoic acid, methyl ester	0.16
2-cyclohexen-1-ol, 2-methyl-5-(1-methylethenyl)-	17.56
5,7,9(11)-androstatriene, 3-hydroxy-17-oxo-	9.51
8,11-octadecadiynoic acid, methyl ester	2.70
9-hexadecenoic acid	9.97
benzene, (1,1-dimethylpropyl)-	20.54
benzene, 1,2,3-trimethyl-	27.69
benzene, 1-ethyl-2-methyl-	34.84
benzene, 1-ethyl-3-methyl-	25.04
benzene, 1-methyl-2-propyl-	20.03
benzene, 1-methyl-4-(1-methylethyl)-	18.55
benzeneacetaldehyde, à-ethyl-	19.03
cis-p-mentha-2,8-dien-1-ol	15.72
cyclopropanedodecanoic acid, 2-octyl-, methyl ester	13.40
dodecane, 5,8-diethyl-	17.17
ethyl iso-allocholate	6.68
heptadecane, 9-hexyl-	14.29
hexadecane, 1,1-bis(dodecyloxy)-	13.23
methyl 10,12-pentacosadiynoate	10.52
methyl 9,11-octadecadiynoate	12.07
nonane	8.00
octadecane, 3-ethyl-5-(2-ethylbutyl)-	14.53
oxiraneoctanoic acid, 3-octyl-, cis-	0.03
undecane	4.23
undecane, 2-methyl-	0.22
1,4-benzenediol, 2,6-bis(1,1-dimethylethyl)-	−2.32
cyclotrisiloxane, hexamethyl-	−20.25
dodecane	−11.13
dodecane, 2,6,10-trimethyl-	−15.86
hexadecane	−12.40
n,n′-pentamethylenebis[s-3-aminopropyl thiosulfuric acid]	−6.64
octadecane, 6-methyl-	−3.24
tetradecane	−18.00
tetradecane, 2,6,10-trimethyl-	−2.68
tridecane	−17.56

4.4. Additional MBC-derived fuel compositional changes

Tables 6–9 show the results of using the fast version of the PARAFAC-based analysis strategy, augmented with ANOVA, with all four fuel data sets described in Section 2.1. Although there are a few non-fuel compounds reported in these tables that are the result of experimental or chromatographic artifacts, such as the appearance of hexamethylcyclotrisiloxane via column bleed, they are left in the tables for the sake of completeness and do not detract from the conclusions drawn. The fast-version results are presented because the more unwieldy results produced using the thorough version of the strategy are not necessary to prove the concept behind the analysis. For the most straightforward example, one need only consult Table 6, which contains the 5% biodiesel analysis results. As described and shown previously, the analysis indicated that there was a relative decrease in several different methyl esters in the B5 sample as they were metabolized by the microorganisms present in the experimental samples. Concurrently, the concentrations of several cyclic compounds were found to increase, suggesting that these compounds were being produced during microbiological growth as either direct or indirect MBC by-products.

The results shown in Tables 7 and 8, together, present an interesting combined assessment of the impact of microbiological growth in petrochemical fuels. As a general trend in the specification F-76 diesel fuel, larger hydrocarbons with low amounts of branching, i.e., C12 and larger straight-chain hydrocarbons, some of which possessing methyl groups, are converted through the presence of microorganisms into smaller chains, more branched chains, and cyclic compounds. Furthermore, although the concentrations of no less than fifteen compounds were identified as increasing or decreasing in the same manner in both the F-76 high-sulfur diesel fuel and the ULSD fuel, the compounds that decreased in both fuels during microbial growth did not include the larger, low-branching hydrocarbons. This indicates that MBC detracts from the chemical compositions of these diesel fuels in differing fashions yet produces similar metabolic products in both. The fact that three of the five components found to increase in the B5 fuel were also found to increase in these petrochemical fuels also indicates a similarity in metabolic products. It was also found that ten of the petrochemically shared fifteen compounds showed a more pronounced

Table 8
"Fast" ANOVA-augmented 1-PARAFAC analysis results obtained from the ULSD data set.

Ultra-low sulfur diesel, or ULSD (petrochemical)	Percent increase (+) or decrease (−)
(e)-3(10)-caren-4-ol	8.93
1,1-cyclobutanedicarboxamide, 2-phenyl-n,n′-bis(1-phenylethyl)-	24.11
10,13-octadecadiynoic acid, methyl ester	22.81
13-heptadecyn-1-ol	5.74
1b,5,5,6a-tetramethyl-octahydro-1-oxa-cyclopropa[a]inden-6-one	3.02
1-decen-4-yne, 2-nitro-	16.33
1-dodecanol, 3,7,11-trimethyl-	23.99
1-octadecanesulphonyl chloride	11.31
2,5-octadecadiynoic acid, methyl ester	6.25
2-cyclohexen-1-ol, 2-methyl-5-(1-methylethenyl)-	20.88
2-piperidinone, n-[4-bromo-n-butyl]-	9.35
3h-cyclodeca[b]furan-2-one, . . .	0.47
4,7-octadecadiynoic acid, methyl ester	3.44
benzene, 1,2,3-trimethyl-	26.04
benzene, 1-ethyl-3-methyl-	29.78
benzene, 1-methyl-4-(1-methylethyl)-	16.13
benzenebutanal	8.67
bicyclo[3.1.0]hexane-6-methanol, 2-hydroxy-1,4,4-trimethyl-	8.33
decane	16.30
dodecane, 5,8-diethyl-	14.54
e-2-octadecadecen-1-ol	14.81
falcarinol	2.75
heptadecane, 9-hexyl-	4.82
hexadecane, 1,1-bis(dodecyloxy)-	18.75
morphinan-4,5-epoxy-3,6-di-ol, 6-[7-nitrobenzofurazan-4-yl]amino-	4.61
naphth[1,2-b]oxirene, decahydro-1a,7-dimethyl-	4.87
n-nonadecanol-1	15.41
octadecane, 6-methyl-	24.13
oxiraneoctanoic acid, 3-octyl-, cis-	6.37
silane, trichlorodocosyl-	16.21
trans-z-à-bisabolene epoxide	3.06
z,z,z-1,4,6,9-nonadecatetraene	6.78
1,2-benzisothiazol-3-amine tbdms	−20.32
1,4-benzenediol, 2,6-bis(1,1-dimethylethyl)-	−18.33
5,7,9(11)-androstatriene, 3-hydroxy-17-oxo-	−14.32
butylaldehyde, 4-benzyloxy-4-[2,2,-dimethyl-4-dioxolanyl]-	−14.59
cyclotrisiloxane, hexamethyl-	−25.42
epi-epoxy-buphanamine	−10.59
epoxybuphanamine	−13.64
n,n′-pentamethylenebis[s-3-aminopropyl thiosulfuric acid]	−13.81
olean-12-ene-3,15,16,21,22,28-hexol, (3á,15à,16à,21á,22à)-	−0.90
perhydroindene-4-carboxylic acid, . . ., methyl ester	−8.31

Table 9
"Fast" ANOVA-augmented 1-PARAFAC analysis results obtained from the FT data set.

Fischer–Tropsch (FT) synthetic diesel fuel	Percent increase (+) or decrease (−)
2,3-dimethyldecane	8.57
decane	12.66
decane, 3-methyl-	10.48
decane, 4-methyl-	12.25
decane, 5-methyl-	12.98
dodecane, 2,6,10-trimethyl-	17.32
dodecane, 2,7,10-trimethyl-	15.77
dodecane, 5,8-diethyl-	2.59
nonane	13.70
nonane, 3-methyl-	16.82
octadecane, 6-methyl-	16.07
undecane	13.54
undecane, 2,6-dimethyl-	14.21
heptadecane	−1.86
hexadecane	−3.19
hexadecane, 3-methyl-	−2.48
n,n′-pentamethylenebis[s-3-aminopropyl thiosulfuric acid]	−2.21
nonadecane	−0.13
octadecane	−1.81
tetradecane	−2.46
tridecane, 3-methyl-	−4.77
tridecane, 4-methyl-	−4.00

percent change in the ULSD fuel than in the F-76 diesel, which suggests that the effects of MBC are more pronounced in ULSD than in high-sulfur diesel fuels.

As expected, the compounds identified in the FT synthetic fuel, before and after MBC, were generally much less complex, with fewer branched alkanes and functional groups, than found in either the petroleum-derived fuels or the B5 fuel (Table 9). This makes some sense, as the Fischer–Tropsch process tends towards products that are not as chemically diverse as those found in the other fuel types [25]. The major impact of MBC on the FT fuel composition was the conversion of straight-chain alkanes to more heavily branched hydrocarbons.

5. Conclusions

An alternative peak selection strategy for quantifying changes between GC–MS data populations was developed based on an interval-oriented one-factor PARAFAC spectral interpretation augmented with an ANOVA-based result-filtering step. Not only is this strategy fundamentally effective, but due to its reliability it is also well-suited to the construction of an automated analysis strategy for use by non-expert fuel analysts. It has been shown to be effective with overly complex and noisy GC–MS data, a situation within which a deconvolution-based peak selection strategy had difficulty functioning properly. This strategy was successfully used to assess four different types of diesel fuel to determine the chemical changes that occur within each of them upon MBC. As no portion of the analysis strategy explicitly relies upon the assessment of fuel analytes, other similarly complex mixed-hydrocarbon analytes and resulting GC–MS data populations could likely be assessed in a similarly thorough fashion. Future work will center upon the implementation of this technique within a self-contained user interface being developed within our laboratory.

Acknowledgement

We thank the Office of Naval Research for their funding and support.

References

[1] G. Liu, L. Wang, H. Qu, H. Shen, X. Zhang, S. Zhang, M. Zhentao, Fuel 86 (2007) 2551.
[2] A.M. Hupp, L.J. Marshall, D.L. Campbell, R.W. Smith, V.L. McGuffin, Anal. Chim. Acta 606 (2008) 159.
[3] R. Fernandez-Varela, J.M. Andrade, S. Muniategui, D. Prada, F. Ramirez-Villalobos, Water Res. 43 (2009) 1015.
[4] S.E. Stein, J. Am. Soc. Mass Spectrom. 10 (1999) 770.
[5] AMDIS: http://chemdata.nist.gov/mass-spc/amdis/ (last accessed 8/3/2010).
[6] O.Y. Begak, A.M. Syroezhko, Russ. J. Appl. Chem. 77 (2004) 653.
[7] R.A. Harshman, M.E. Lundy, Comput. Stat. Data Anal. 18 (1994) 39.
[8] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[9] M.L. Nahorniak, G.A. Cooper, Y.-C. Kim, K.S. Booksh, Analyst 130 (2005) 85.
[10] A. Niazi, A. Yazdanipour, Pharm. Chem. J. 41 (2007) 170.
[11] S.-H. Zhu, H.-L. Wu, A.-L. Xia, Q.-J. Han, Y. Zhang, R.-Q. Yu, Talanta 74 (2008) 1579.
[12] K. Faber, A. Lorber, B.R. Kowalski, J. Chemom. 11 (1997) 95.
[13] C.G. Fraga, J. Chromatogr. A 1019 (2003) 31.
[14] C.G. Fraga, C.A. Corley, J. Chromatogr. A 1096 (2005) 40.
[15] J.L. Hope, A.E. Sinha, B.J. Prazen, R.E. Synovec, J. Chromatogr. A 1086 (2005) 185.
[16] J.C. Hoggard, R.E. Synovec, Anal. Chem. 80 (2008) 6677.
[17] J.C. Hoggard, W.C. Siegler, R.E. Synovec, J. Chemom. 23 (2009) 421.
[18] D.M. Stamper, R.E. Morris, Poster Presentation at the 110th General Meeting of the American Society for Microbiology, May, San Diego, CA, 2010.
[19] NIST Standard Reference Database 1A: http://www.nist.gov/ts/msd/srd/nist1.cfm (last accessed 8/3/2010).
[20] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 80 (1998) 17.
[21] R.C. Prince, C. Haitmanek, C.C. Lee, Chemosphere 71 (2008) 1446.
[22] R.K. Murray, D.K. Granner, P.A. Mayes, V.W. Rodwell, Harper's Illustrated Biochemistry, 26th ed., Lange Medical Books/McGraw-Hill: Medical Publishing Division, 2003, p. 122, 180.
[23] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley and Sons Ltd., 2003, p. 23.
[24] M.T. Cantwell, S.E.G. Porter, S.C. Rutan, J. Chemom. 21 (2007) 335.
[25] G. Jacobs, K. Chaudhari, D. Sparks, Y. Zhang, B. Shi, R. Spicer, T.K. Das, J. Li, B.H. Davis, Fuel 82 (2003) 1251.