Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Chromatography A, 1096 (2005) 156–164

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Classification of highly similar crude oils using data sets from comprehensive two-dimensional gas chromatography and multivariate techniques

V.G. van Mispelaar[a,*], A.K. Smilde[b,c], O.E. de Noord[a], J. Blomberg[a], P.J. Schoenmakers[c]

[a] *Shell Global Solutions International B.V., Analytical Problem Solving Amsterdam, P.O. Box 38000, 1030 BN Amsterdam, The Netherlands*
[b] *TNO Quality of Life, Zeist, The Netherlands*
[c] *University of Amsterdam, Amsterdam, The Netherlands*

Available online 19 October 2005

Abstract

Comprehensive two-dimensional gas chromatography (GC × GC) has proven to be an extremely powerful separation technique for the analysis of complex volatile mixtures. This separation power can be used to discriminate between highly similar samples. In this article we will describe the use of GC × GC for the discrimination of crude oils from different reservoirs within one oil field. These highly complex chromatograms contain about 6000 individual, quantified components. Unfortunately, small differences in most of these 6000 components characterize the difference between these reservoirs. For this reason, multivariate-analysis (MVA) techniques are required for finding chemical profiles describing the differences between the reservoirs. Unfortunately, such methods cannot discern between 'informative variables', or peaks describing differences between samples, and 'uninformative variables', or peaks not describing relevant differences. For this reason, variable selection techniques are required. A selection based on information between duplicate measurements was used. With this information, 292 peaks were used for building a discrimination model. Validation was performed using the ratio of the sum of distances between groups and the sum of distances within groups. This step resulted in the detection of an outlier, which could be traced to a production problem, which could be explained retrospectively.

Keywords: GC×GC; Crude oil characterization; Discrimination; Clustering; Validation

1. Introduction

Chemists working in (gas) chromatography are continuously faced with improved instrumentation and techniques. Developments in injection techniques facilitate the injection of large volumes and 'dirty' samples, while selective detection allows detection of components at low levels. Moreover, developments in electronics, such as flow control, strongly improve the repeatability and reproducibility of the technique.

All these developments have resulted in a dramatically enhanced robustness of (gas) chromatographic methods. They also create the possibility to analyze large numbers of samples in a more-or-less automated way, facilitating other types of applications, such as high-throughput analysis and metabolism studies (metabolomics).

Instrumental advances also affect the applicability of comprehensive two-dimensional gas chromatography (GC × GC). With this technique, highly complex semi-volatile mixtures can be analyzed in unsurpassed detail. GC × GC can separate complex samples into thousands of individual components. Most examples in the literature concern a single or a few samples. However, the comparison of a series of GC × GC chromatograms can yield very valuable information as well. Especially for highly similar samples, high-resolution techniques are essential to reveal minute differences.

Large data sets require other processing approaches than conventional chromatograms. If there is no prior information regarding components of interest, the traditional approach of quantifying all components present and comparing them univariately is clearly not an attractive strategy. Multivariate-analysis (MVA) techniques provide better options for processing such large data sets. Such an approach is already adopted in, for example, the field of metabolomics [1,2]. By comparing two (or more) groups of subjects (e.g. sick versus healthy, treated versus untreated), valuable information on metabolic differences between these groups can be obtained. However, this information can only be attained when the number of objects is sufficiently large to eliminate natural variation between the sub-

[*] Corresponding author. Tel.: +31 20 630 3601; fax: +31 20 630 2911.
E-mail address: v.vanmispelaar@shell.com (V.G. van Mispelaar).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.09.063

jects. This approach is not restricted to systems biology. It is also is applicable to other highly complex mixtures, such as crude oil. The chemical composition of crude oil is determined by its origin and geochemical history. Both chemical composition and boiling-point range can vary widely between different oil fields. However, different reservoirs within one field have a similar origin and a highly similar geochemical history, which can result in minute differences in chemical composition.

Crude oil can contain hydrocarbons from C_4 up to C_{100} or even higher, and the number of theoretical isomers is stunningly large. Techniques such as GC × GC and GC × GC–MS are by far not sufficient to reveal the full complexity of this class of mixtures. The number of components that can be separated and identified using these techniques is nonetheless impressive.

From a chemometric point of view, these chromatograms are highly interesting. Each object (or sample) is described by a large number of variables (or peaks). Classification of these objects according to their origin can be achieved using discriminant-analysis (DA) methods. Such methods try to find profiles of variables in the data that differentiate between groups of objects. A priori information (which object belongs to which group) is required. In many cases this information seems obvious. For example, patients are healthy or sick. However, this information is not necessarily correct. In the example used, patients may not be diagnosed correctly or they may be suffering from other disorders. Incorporating incorrect information into these so-called “supervised techniques” will clearly lead to erroneous models. On the other hand, exploratory techniques, such as principal component analysis, are not beneficial if the data contains a high number of uninformative variables.

Therefore, a combination of supervised techniques, for the discovery of discriminating variables, and unsupervised techniques, for finding natural clusters in data, is potentially very strong.

In this study, we will apply GC × GC to a set of crude-oil samples from three reservoirs within an oil field. Since no prior knowledge was available on the chemical components that would discriminate between the three fields, as many components as possible needed to be separated and quantified. Multivariate-analysis techniques facilitated the recovery of discriminating components or component profiles.

2. Theory

2.1. GC × GC

One of the greatest and most significant advances for the characterization of complex mixtures of volatile compounds has been the advent of comprehensive two-dimensional gas chromatography. This technique was pioneered and advocated by the late Phillips and co-workers [3,4]. Two different GC columns are used in GC × GC. The fist-dimension column is (usually) a conventional capillary GC column, with a typical internal diameter of 250 μm. Most commonly, this column contains a non-polar stationary phase, so that it separates components largely based on their vapor pressures (boiling points). The second-dimension column is considerably smaller (smaller diameter, shorter length) than the first-dimension column, so that separations in the second dimension are essentially much faster. The stationary phase is selected such that this column separates on properties other than volatility, such as molecular shape or polarity. The two columns are coupled using a so-called modulator. This device facilitates the continuous accumulation, refocusing and injection of small portions of the first-column effluent into the second-dimension column. With each modulation, a new second-dimension chromatogram is started. The detector, which is positioned at the end of the second-dimension column, records these fast chromatograms. At the end of a chromatographic run, the chromatogram contains many of these fast separations in series. After ‘demodulation’ [5], a two-dimensional chromatogram is obtained, which is usually represented by a colour or contour plot.

In many applications, GC × GC has proven to be an excellent technique for the separation of very complex samples, such as petrochemical products [6–8], essential oils [9,10], fatty acids [11,12], doping control [13], flavour analysis [14], residue analysis [15], and cigarette smoke [16]. The combination of GC × GC and MVA techniques is described in various references with the aim of deconvolution [17] and enhancing the detection limit [18]. Johnson et al. describe the use of a high-speed GC × GC for pattern recognition of jet fuels [19].

2.2. Data analysis

Many analytical techniques exist that can generate large data sets. The human mind is only capable of interpreting data in three-dimensions. Visualization of higher-dimensionality data requires reduction techniques. Fortunately, multivariate analysis offers various approaches to reduce the data dimensionality.

Classification and clustering problems can be solved using two types of techniques. Exploratory methods extract (natural) patterns in the data. Supervised classification techniques use prior information (which objects belong to which groups) to find differences (or similarities) between groups of samples.

2.2.1. Exploratory methods

2.2.1.1. PCA. The most commonly encountered exploratory method is principal-component analysis (PCA). In PCA, the original variables are replaced by a (strongly) reduced number of uncorrelated (orthogonal) variables, called the principal components.

Mathematically:

$$X = T \times P^T + E \tag{1}$$

where X: original data set containing n (samples) × p (variables); T: scores n (samples) × F (principal components); P^{T}: transposed loadings containing F (principal components) × p (variables); E: residuals, variation not explained by model

The principal components are constructed in such a way, that the first principal component (PC1) represents the main source of variation in the original data set. The second PC is orthogonal to the first and represents the maximum variance not explained in PC1. The third PC again is orthogonal to the first two PC's etc. Each PC is a linear combination of the original variables. The contribution of each variable is expressed in the principal-component loadings.

The number of PCs gives an indication of the model complexity. If the data are highly correlated, a few PCs will be sufficient to reproduce the original data. A way of presenting the data from this technique is the score plot. Related objects (belonging to the same groups) have similar scores and will consequently tend to cluster.

2.2.1.2. Projection pursuit. Another unsupervised projection technique is projection pursuit (PP) [20]. Unlike PCA, the main objective of which is to explain variance in the data, PP searches interesting low-dimensional linear projections in the data. This is achieved by optimizing the projection index (PI), which can be regarded as an objective function. In literature, several projection indices have been described [20].

2.2.2. Supervised techniques

Discriminant-analysis [21,22] methods can be applied if attention is focused on differences between known groups of samples. The technique is based on the assumption that samples of the same group are more similar than samples belonging to different groups. The goal of DA is to find and identify structures in the original data, which show large differences between the group means. This process requires a priori knowledge on which samples belong to the same group.

Discriminant analysis has been used for a wide variety of problems in analytical chemistry. For example, the differentiation of coffee [23,24], wine [25–27], and many other types of samples has been described.

Many discriminant methods have been described in the literature. Both Fischer's linear discriminant analysis (FLDA) [28] and quadratic discriminant analysis (QDA) [28] can be used in cases where the number of objects (greatly) exceeds the number of variables. In situations where the number of variables exceeds the number of objects, PCA and partial least squares (PLS) are used to reduce the dimensionality of the data. The principal components or latent variables are then subjected to linear discriminant analysis. These techniques are described in the literature as partial-least-squares discrimination analysis (PLSDA) [29] and principal-component discriminant analysis (PCDA) [30]. They have been used successfully in various types of applications. Regularized discriminant analysis (RDA) [31] has been proposed for data sets where the number of variables only slightly exceeds the number of objects.

2.2.2.1. PCDA. Discriminant analysis of data containing more variables than objects can be preceded by principal-component analysis to reduce the number of variables. The projections (scores) of the samples on the principal components are used as a starting point for FLDA. Graphical representation of both the objects (in a score plot) and the discriminant loadings provides valuable information on relations between objects and on important variables in the data set.

2.2.3. Validation

There are several ways to validate a (discrimination) model. In cross validation one or several objects are excluded, model is created using the remaining objects, and the group membership of the excluded sample is predicted. A well-balanced model results in a minimal number of false assignments.

Another method to assess the validity of a model is by permutation. In this process, the effects of the random assignment of objects to groups are examined.

Fig. 1 gives a graphical representation of a hypothetical data set.

In this figure, twelve objects are described by two variables, X_1 and X_2, located in three groups. Suitable classification of these objects would lead to three dense populations, whereas the distances between the populations should be large. The 'within-

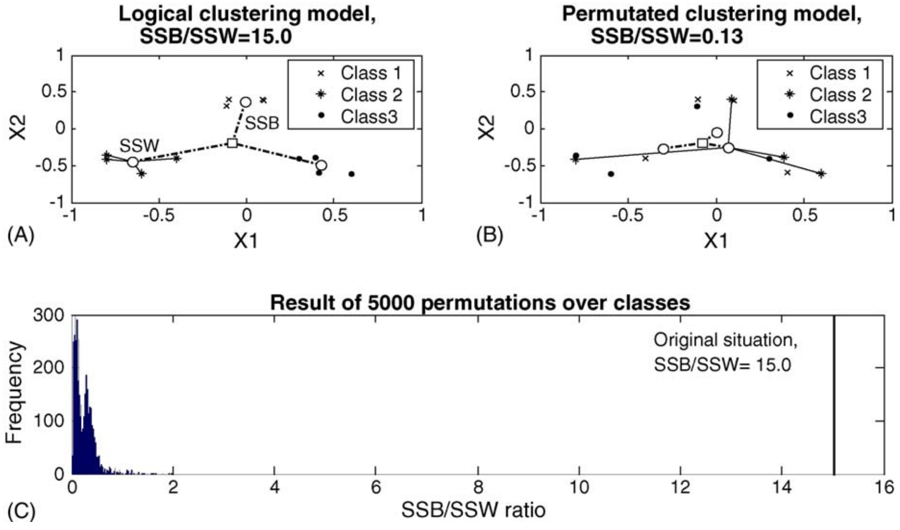

Fig. 1. Explanation of sum-of-squares within and sum-of-squares between groups.

group distance' gives a measure for the density of the clusters and can be obtained by calculating the distances for each object to the centre of its group. The 'between-group distance' can be used as a measure for the separation of the three clusters and is obtained by calculating the distance between the group centres. The ratio of the 'sum of distances between groups' and the 'sum of distances within' groups should be maximal for proper clustering. Since the 'sum-of-squares' is used in the calculation, we will refer to SSB and SSW for sum-of-squares distances between and within the groups, respectively. The initial situation in Fig. 1A results in an SSB/SSW ratio of 15.0. In a permutation process, objects are assigned randomly to one of the three groups. The result of a first (random) permutation is shown in Fig. 1B. The sum of distances within the clusters increases significantly, while the sum of distances between the groups changes only slightly. This has a dramatic effect on the ratio SSB/SSW (0.1). Repeating this permutation process many times results in equally many SSB/SSW ratios. A histogram of all these results is presented in Fig. 1C. Most of the random permutations result in SSB/SSW values between 0 and 1. The original situation, with a SSB/SSW ratio of 15.0 is clearly the best classification of the data.

The above calculations can be described mathematically [32] for the between-group distance:

$$\mathrm{SSB} = \sum_{i=1}^{g} m_i \times (\bar{x}_i - \bar{x})^2 \qquad (2)$$

For the within-group distance:

$$\mathrm{SSW} = \sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2 \qquad (3)$$

where g: number of groups; m_i: number of objects for group i; x_{ij}: object i of group j.; $\bar{x}_i$: mean of group i.; $\bar{x}$: overall mean of $\bar{x}_i$

3. Experimental

3.1. Instrumentation

The samples were analyzed using an Agilent 6890 GC, (Wilmington, DE, USA), equipped with a CTC CombiPAL autosampling unit (CTC Analytics, Zwingen, Switzerland), and a CIS-4 Programmed-Temperature-Vaporization (PTV) injector (Gerstel, Mulheim an der Rühr, Germany). This system was retrofitted with a Zoex KT2003 thermal modulator and equipped with a second dimension-column oven (Zoex, Lincoln, NE, USA), enabling independent second-dimension column heating. The column set consisted of a 10 m length × 0.25 mm internal diameter × 0.25 μm film thickness DB1 column (J&W Scientific, Folsom, CA, USA) in the first dimension and a 2 m length × 0.1 mm internal diameter × 0.1 μm film thickness BPX50 column (SGE, Ringwood, Australia) in the second dimension. The modulation was performed in a 1.6 m × 0.1 mm DPTMS deactivated fused silica capillary (BGB Analytik, Anwil, Switzerland). A fused-silica capillary of the same material with a length of approximately 50 cm was used to connect the second-dimension column to the flame-ionization detector (FID). Columns were coupled with custom-made press-fits (Techrom, Purmerend, The Netherlands). The carrier gas was Helium at a constant head pressure of 250 kPa, resulting in a column flow of approximately 1 mL/min at 40 °C. The temperature for the first-dimension column oven was programmed from 40 °C (5 min isothermal) at a rate of 2.5 °C/min to 300 °C (20 min isothermal), followed by a negative ramp of 13 °C/min to 40 °C (10 min isothermal).

The PTV injector was programmed from 40 °C to 250 °C (5 min isothermal) with a ramp of 12 °C/s. These conditions resulted in a selective discrimination from C_{30} hydrocarbon upwards. The advantages of this approach are introduction of a minimum of residual material and reducing the maximum oven temperature, which reduces column degradation. These steps result in an increased chromatographic stability in terms of retention time. The modulation time was 7.5 s and the hot-pulse duration was 300 ms. Both the hot pulse of the release jet and the secondary oven were operated at an offset of 50 °C above oven temperature. Liquid-nitrogen-cooled nitrogen gas was used as modulating agent at a flow of ∼17 L/min. A Zoex auto-fill unit was used to enable continuous operation.

3.2. Instrument control and data processing

Instrument control and data acquisition were achieved with EZ-Chrom elite (v2.61, Scientific Software Europe, Willemstad, the Netherlands). Data were collected at 100 Hz to obtain sufficient data points across a peak. Chromatograms were exported to the Chromatography Data Format (CDF, or AIA level 2).

Data handling was performed in Matlab R14 (The Mathworks, Natick, MA, USA) running on a Compaq EVO W6000 computer equipped with 1 Gb of RAM. Data-handling routines were developed in-house. In addition, the NetCDF toolbox (US Geological Survey, Woods Hole, MA, USA) was used.

3.3. Samples

A set of 14 different oil samples, originating from one oil field, was selected by Shell International Exploration and Production (SIEP, Rijswijk, The Netherlands). The samples were divided in the three subclasses A, B and C, referring to the reservoirs within the original oil field.

The samples were diluted ten-fold in cyclohexane (p.a. quality, Merck, Darmstadt, Germany) containing 0.1% (w/w) 1,2-dichlorobenzene (p.a. quality, Merck) as an internal standard.

All samples were analyzed in duplicate. One sample was analyzed in five-fold. In the sequence two blanks and an alkane mixture, containing C_5–C_{42} hydrocarbons in CS_2, were included.

4. Results and discussion

Samples were analyzed in a sequence in order to reduce retention variations. The negative ramp in the oven program was used to obtain a highly repeatable temperature program, thereby reducing retention-time shifts. The alkane mixture was used to *"spline"* the data. In this process, *n*-alkanes were shifted

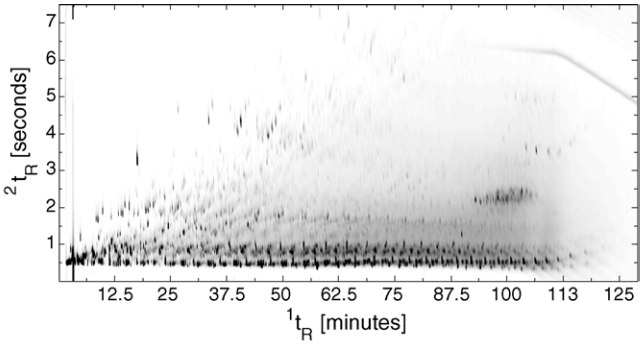

Fig. 2. Typical GC × GC chromatogram of a crude oil.

to obtain constant second-dimension (relative) retention times. The peak positions for a homologous series of *n*-alkanes were used to create a piecewise-linear shift function. This function was subsequently applied to all samples in the sequence.

Fig. 2 shows the two-dimensional chromatogram of a typical crude-oil sample.

Since the data were acquired directly from the FID, a series of second-dimension chromatograms was registered. Integration of this signal therefore resulted in integrated second-dimension chromatograms. However, the modulation process typically resulted in three or four modulations across a first-dimension peak. The total peak area of a certain component is the sum of areas in the successive modulations. The automated summation in the first dimension direction was performed with an algorithm developed in-house. This algorithm goes through the data matrix in which integrated peak positions are put. After location of a peak apex, it runs through a predefined path in the direction of the first dimension axis. This path accounts for a certain deviation in 2nd dimension position, due to concentration effects (resulting in slightly fronting peaks and a higher second dimension retention time) as well as for temperature effects (each successive modulation is injected in a slightly higher 2nd dimension oven temperature). If only one peak is found in the search path, the response for the peak is the sum of the peak areas. If multiple peaks are found, the first derivative of the peak profile is used to determine valleys between the peaks. Wrap-around is circumvented by continuing the search path beyond the limits of the GC × GC chromatogram. In this step, the number of peak positions found was typically reduced by a factor of 2.5, indicating an average of 2.5 modulations over a first-dimension peak.

4.1. Pre-processing

Typical crudes can contain about 6000 individual integrated peaks. This number includes peaks eluting in the isothermal region of the chromatogram and components that are not interesting for quantitative analysis due to the selective discrimination of the PTV. Components eluting from a first-dimension retention time of 106 min upwards were not quantitatively transferred from the injector to the column and therefore eliminated.

4.1.1. Alignment

Unfortunately, chromatographic techniques suffer from retention shifts. This results in inconsistent retention times within a series of samples. Since MVA techniques are unable to deal with shifting peaks, alignment is a required pre-processing step. The alignment algorithm for this study aims to eliminate retention-time shifts for integrated peaks. The algorithm initializes by positioning all found (combined) peak apices at their retention coordinate in an imaginary data matrix. Subsequently, the algorithm walks through the first data matrix for finding peaks. If a peak is found, all other samples are checked whether they also contain a peak in the same region. A maximum shift of 15 datapoints in the second dimension axis and a shift of 1 datapoint in the first dimension direction is used to eliminate differences in retention. If no peak was found in the selected region, the peak arc was set at zero. Found peaks in the data matrix were replaced by a zero entry. After completion of the first data matrix, all other matrices were treated in a similar way. In the end, no more peak areas were found in any of the data matrices. A final check was performed to compare the total peak area before and after alignment. An alignment routine developed in-house was used to eliminate small variations in peak-apex locations. The maximum allowed shift was one data point in the first dimension (=7.5 s) and 15 points in the second dimension (=0.15 s). Fig. 3 shows the position of the aligned peaks.

After cut-off and alignment, a selection of 3904 peaks remained. The resulting dataset contained 30 × 3904 (objects or samples × variables or peaks). Such well-described data should (at least theoretically) be very suitable for multivariate-analysis techniques. However, the PCDA result after ‘mean-centering’ was disappointing (Fig. 4).

A good classification model should form dense, separated clusters. In our initial situation PCDA resulted in overlapping clusters, indicating no separation between groups. The SSB/SSW plot (Fig. 5) also turned out to be highly unsatisfactory. The proposed classification turned out to be no better than a random classification.

This observation can partly be explained by ‘over-fitting’. Since each object is described by 3904 variables, the number of objects should be much larger than 30 to obtain proper classification results. This seems trivial, but it is a very important problem within the MVA field. Many analytical techniques are able to provide highly detailed information, resulting in large

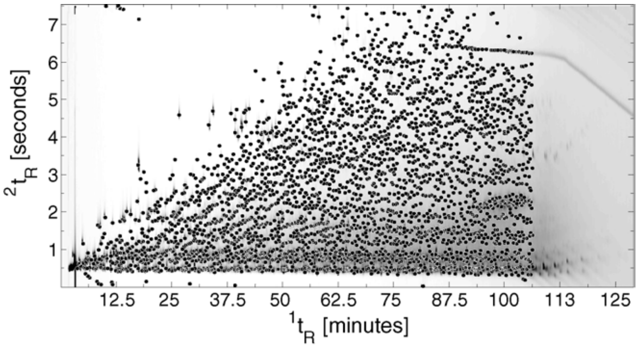

Fig. 3. Location of peaks after alignment.

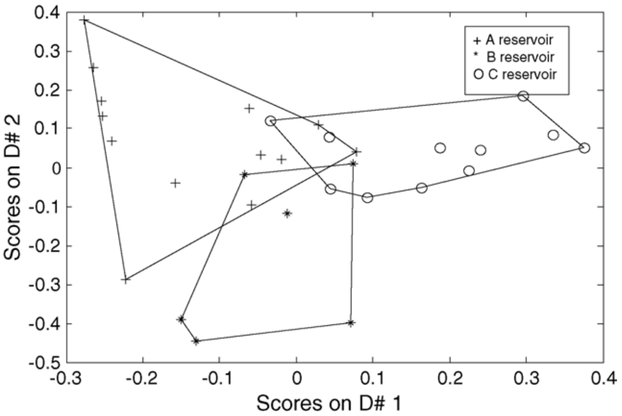

Fig. 4. PCDA of 3904 aligned peaks in 30 objects.

sets of data. The number of available samples usually does not increase accordingly. Even dimension reduction using PCA was insufficient to abstract sufficient relevant information, despite the captured variance in 10 PC's being 83%.

The second problem is the presence of uninformative data, which is irrelevant for the differentiation between reservoirs. Many peaks are indeed of little or no relevance. They contribute hardly or not to the desired discrimination model. The other source of irrelevant data is the integration process. Integration of highly complex chromatographic signals inevitably results in errors. Baseline-separated peaks can be quantified very accurately; convoluted peaks are much-more difficult to integrate. In the case of crude oil, certain regions of the chromatogram do not contain any baseline, due to the continuous elution of components. Quantification of such a signal obviously does not yield relevant data, since the integration errors obscure relevant information.

However, these irrelevant data are included when building the discrimination model and performing the SSB/SSW calculation. Distinction between informative and uninformative peaks can be achieved by variable selection.

4.1.2. Variable selection

Variable selection is commonly performed to (strongly) reduce the number of variables in a dataset. However, many variable-selection strategies can be considered to be supervised, i.e. variables are identified which support a certain group structure. Since supervised selection routines aim at finding components supporting the proposed classification structure. However, a different classification structure will also lead to a selection of components. These routines are therefore solely dependent on the classification structure. Even in a dataset containing random numbers, supervised classification is capable of selecting a number of (random) variables that would support the classification.

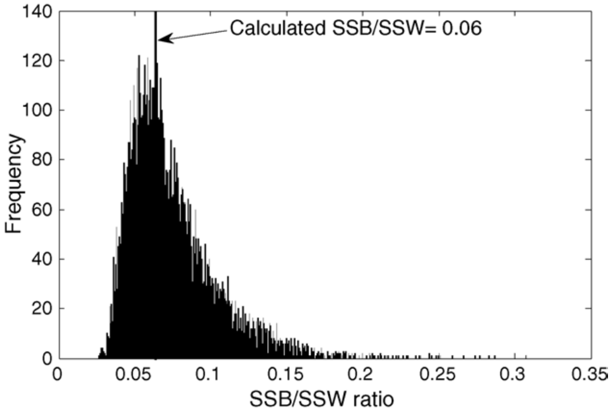

Fig. 5. SSB/SSW distribution of 1000 random permutations.

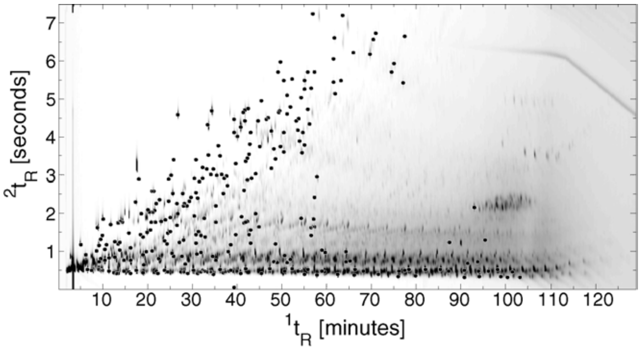

Fig. 6. Peaks selected on RSD between duplicates.

Unsupervised variable selection seems, therefore, to be a more appropriate choice. A suitable criterion can be established by using the information gained from duplicate measurements. Well-separated peaks in duplicate measurements show small relative standard deviations (RSD). All of the 14 samples were analysed in duplicate, yielding 3904 peaks. Reliably quantified peaks should have a small relative standard deviation between duplicate measurements. This should be the case for all of the 14 duplicate measurements. In addition, the average RSD of the 14 samples should be under a certain threshold. Our variable-selection technique selects only those peaks with an average RSD under 10%. In our situation, there were 14 samples and thus as many RSD values for each of the 3904 peaks. A selected peak should have a small RSD value for each of the 14 samples. The average RSD value for any component over all the 14 samples should be as low as possible. Also, the standard deviation between the 14 RSD values should be minimized, excluding variables for which one of the samples has a large RSD, while other objects have a small RSD value.

By restricting the average relative RSD between duplicate measurements (for each of the 3904 peaks) to 10%, 292 variables were selected. Fig. 6 shows the positions of the selected peaks.

Subsequent PCDA revealed clustering according to the reservoir origin. Inspection of the DA loadings did not reveal any specific 'biomarker components' that could be used to discriminate between reservoirs. Differences between the three reservoirs were the result of many small differences between the 292 selected peaks.

4.1.3. Manual selection

Samples from the different reservoirs could not be discriminated based on one or a few components (so-called biomarkers). Rather the differences in all of the included peaks had

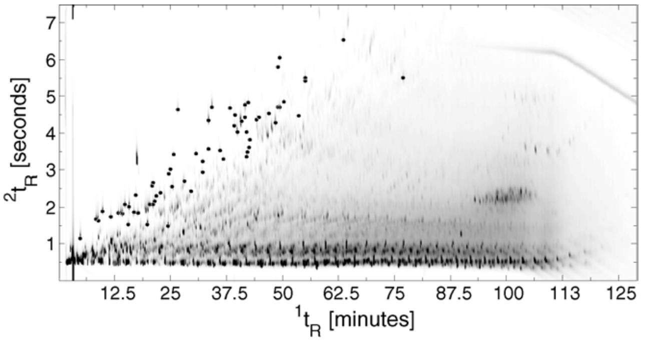

Fig. 7. Position of 65 manual selected peaks.

to be considered. Therefore, verification of the groups had to be performed using unsupervised MVA techniques applied to a small subset of the data. To this end, a selection of 65 baseline-separated peaks was manually extracted from the chromatograms. Fig. 7 shows the peak positions.

The resulting dataset was significantly better defined, having the dimension of 30 × 65 (samples × peaks). Before data-analysis, mean centering was applied as a pre-processing technique. Fig. 8 shows the result of principal-component analysis.

With only two PC's, 96.9% of the variance was captured. Samples in all groups (A, B and C) formed dense clusters, implying a high similarity between the members of each group. However, both duplicates of sample 4S94A seem to be very different from the other A-group members. These samples have therefore to be considered as outliers. Based on these results, the two samples are likely to belong to a different group (e.g. originate from a different reservoir). Calculation of SSB/SSW values can numerically support the outlier hypothesis.

The distance between the two duplicates of 4S94A can be explained by the small percentage of variation in PC2. Small differences between the samples are blown out of proportion. Projection pursuit yielded a somewhat improved clustering results, as shown in Fig. 9.

The observation that two samples are not classified correctly obviously has severe implications for discriminant analysis. Grouping/clustering of samples of incorrect origin evidently results in the calculation of incorrect DA-loadings and scores.

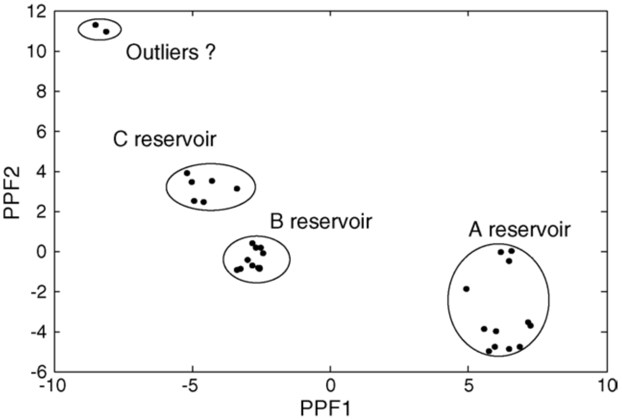

Fig. 9. Projection pursuit after mean centering.

There are a number of possible solutions for this problem. The first is to simply remove the two samples from the dataset before the PCDA step. Trying to fit the samples into one of the two other groups may also be a possible solution. The third option is to define a new group in which the two duplicates of the samples are included.

These three options were all investigated. Results can be found in Figs. 10 and 11.

The SSB/SSW ratios were dramatically improved, while the discriminant analysis resulted in much denser clusters. Based on these results, sample 4S94A is best described as a new group, since this hypothesis results in the best (highest) SSB/SSW ratio.

The suppliers of the samples gave the ultimate proof for the hypothesis that sample 4S94A was different from the other A-group members. This specific sample was taken during a pipeline leakage. Instead of producing A-product, a mixture of A and C was produced. The fourth option in which a separate class is defined for the two samples in Figs. 10 and 11d describes this situation best. However, this conclusion is not in line with the PCA results. Since PCA scores are a linear combination, mixtures of the groups would fall in-between the pure groups. This could be (only partly) explained by the small number of components (65 out of 3904), which may not be representative for the entire sample.

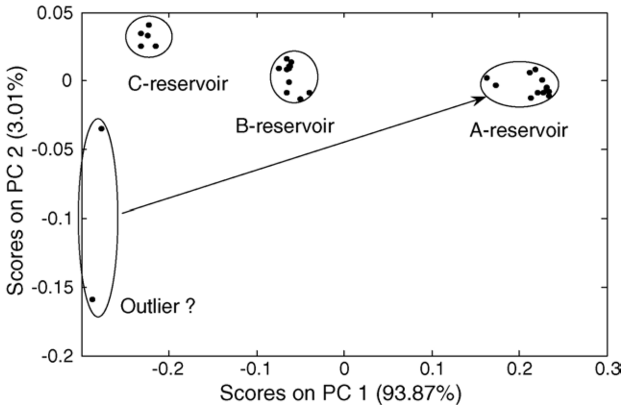

Fig. 8. PCA after mean-centering of 65 manually selected peaks.

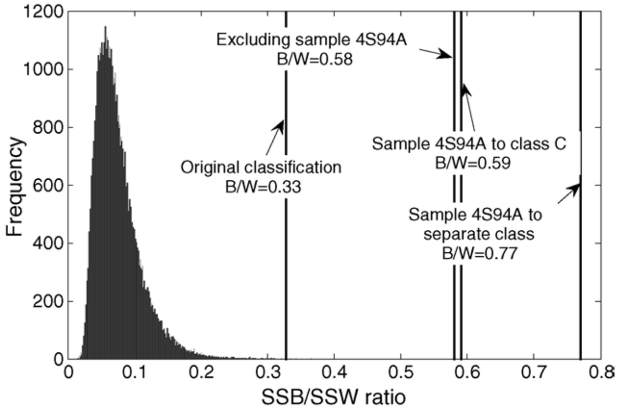

Fig. 10. SSB/SSW results of 1000 random permutations.

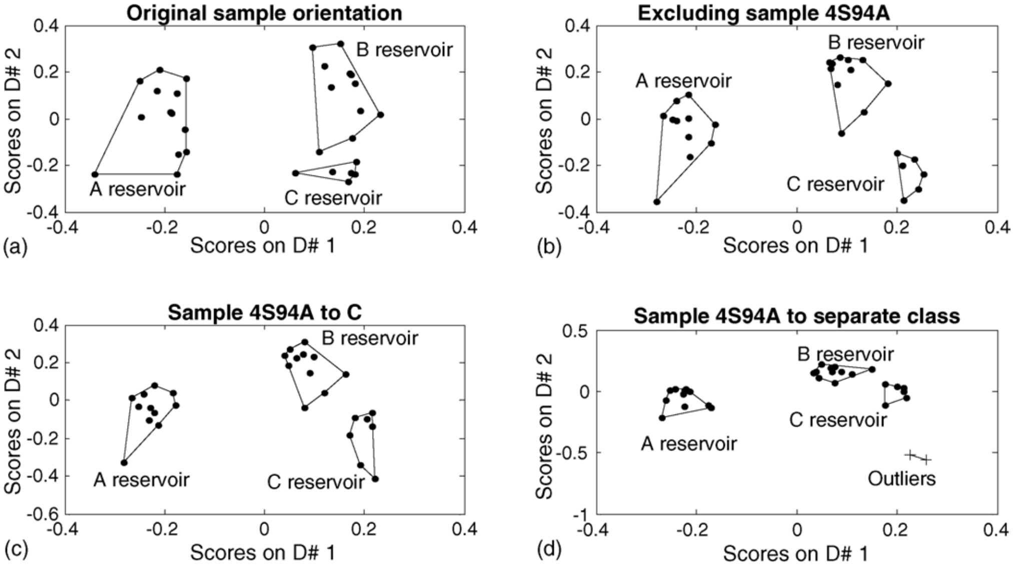

Fig. 11. PCDA results of different scenarios.

5. Conclusions

Comprehensive two-dimensional gas chromatography again turns out to be highly suitable for measuring small differences between complex samples. This has already been demonstrated in various examples in the literature. However, improved modulation techniques (e.g. cryogenic modulation) lead to a drastically improved stability of retention times. This facilitates the comparison of large series of samples and the use of sophisticated (multi-variate) data-analysis methods.

The measured samples, consisting of crude oils from three reservoirs within one oil field, were highly similar. The necessary pre-processing techniques, such as integration and alignment, resulted in 3904 peaks found in all 30 samples. However, using discriminant analysis on this dataset, we were unable to calculate discriminant scores based on which the samples could be separated according to their origin. Variable selection turned out to be essential to eliminate the problem of over-determination of the data matrix.

Selection of variables based on the average relative standard deviation between duplicate measurements, with an upper limit to 10%, resulted in a reduction to 292 variables (peaks). When this data was subjected to PCDA, clusters separating the reservoirs appeared.

Verification of the groups with PCA and projection pursuit resulted in the discovery of an outlier. Feeding this information into PCDA did improve the results dramatically. The validation with the ratio of SSB/SSW (sum-of-squares-between-groups and sum-of-squares-within-groups) proved unambiguously that the proposed classification was superior to the original classification. This supported the hypothesis that the initial classification structure contained an incorrect entry.

The outlier in the dataset could be explained retrospectively from production problems.

Acknowledgement

The authors would like to acknowledge Florian Wülfert for many stimulating discussions on classification techniques.

References

[1] O. Fiehn, Plant Mol. Biol. 48 (1) (2002) 155.
[2] G.G. Harrigan, R. Goodacre (Eds.), Metabolic Profiling: Its Role to Biomarker Discovery and Gene Function Analysis, Kluwer Academic Publishing, Boston, USA, 2003.
[3] Z. Liu, J.B. Phillips, J. Chromatogr. Sci. 29 (1991) 227.
[4] J.B. Phillips, J. Beens, J. Chromatogr. A. 856 (1999) 331.
[5] J.B. Phillips, D. Liu, J. Pawliszyn, Anal. Chem. 57 (1985) 2779.
[6] J. Blomberg, P.J. Schoenmakers, J. Beens, R. Tijssen, J. High Resol. Chromatogr. 20 (1997) 539.
[7] G.S. Frysinger, R.B. Gaines, J. High Resol. Chromatogr. 23 (2000) 197.
[8] J. Beens, J. Blomberg, P.J. Schoenmakers, J. High Resol. Chromatogr. 23 (2000) 182.
[9] P. Marriott, R. Shellie, J. Fergeus, P. Morrison, Flavour Fragr. J. 15 (2000) 225.
[10] J.-M.D. Dimandja, S.B. Stanfill, J. Grainger, D.G. Patterson Jr., J. High Resol. Chromatogr. 23 (2000) 208.
[11] H.-J. de Geus, I. Aidos, J. de Boer, J.B. Luten, U.A.Th. Brinkman, J. Chromatogr. A 910 (2001) 95.
[12] L. Mondello, A. Casilli, P.Q. Tranchida, P. Dugo, G. Dugo, J. Chromatogr. A. 1019 (2003) 187.
[13] A.J. Kueh, P.J. Marriot, P.M. Wynne, J.H. Vine, J. Chromatogr. A. 1000 (2003) 109.
[14] M. Adachour, J. Beens, R.J.J. Vreuls, A.M. Batenburg, E.A.E. Rosing, U.A.Th. Brinkman, Chromatographia 55 (2002) 361.
[15] J. Dallüge, M. van Rijn, J. Beens, R.J.J. Vreuls, U.A.Th. Brinkman, J. Chromatogr. A. 965 (2002) 207.
[16] J. Dallüge, L.L.P. van Stee, X. Xu, J. Williams, J. Beens, R.J.J. Vreuls, U.A.Th. Brinkman, J. Chromatogr. A. 974 (2002) 169.
[17] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.
[18] C.A. Bruckner, B.J. Prazen, R.E. Synovec, J. High Resol. Chromatogr. 23 (2000) 215.
[19] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 68 (2002) 225.
[20] M. Daszykowski, B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 65 (2003) 97.

[21] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics. Part A, Elsevier, Amsterdam, 1997.

[22] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics. Part B, Elsevier, Amsterdam, 1998.

[23] K.A. Anderson, B.W. Smith, J. Agric. Food Chem. 50 (2002) 2068.

[24] A.J. Charlton, W.H.H. Farrington, P. Brereton, J. Agric. Food Chem. 50 (2002) 3098.

[25] J.T.W.E. Vogels, A.C. Tas, F. van den Berg, J. van der Greef, Chemom. Intell. Lab. Syst. 21 (1993) 249.

[26] M.A. Bresia, V. Caldarola, A. de Giglio, D. Benedetti, F.P. Fanizzi, A. Sacco, Anal. Chem. Acta 458 (2002) 177.

[27] W.-O. Kwan, B. Kowalski, J. Agric. Food Chem. 28 (1980) 356.

[28] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley and Sons, New York, USA, 1992.

[29] M. Barker, W. Raynes, J. Chemometrics 17 (2003) 166.

[30] R. Hoogerbrugge, S.J. Willig, P.G. Kistemaker, Anal. Chem. 55 (1983) 1710.

[31] J.H. Friedman, J. Am. Stat. Assoc. 84 (2003) 165.

[32] W.J. Krzanowski, Principles of Multivariate Analysis, A User’s Perspective, Oxford Science Publications, Oxford, UK, 1988.