9 A Multivariate Approach to Oil Hydrocarbon Fingerprinting and Spill Source Identification

Jan H. Christensen[#] and Giorgio Tomasi*

[#]Department of Natural Sciences, Royal Veterinary and Agricultural University
*Department of Food Science, Royal Veterinary and Agricultural University

9.1 Introduction

Oil hydrocarbon fingerprinting was originally developed by geochemists in the petroleum industry to understand and track the source of crude oils and natural gases. In environmental forensics, methods basically similar to those of petroleum geochemistry have been applied to defensibly determine oil source(s) (Wang et al., 1999, 2002; Stout et al., 2001; Daling et al., 2002; Christensen et al., 2004, 2005d), distinguish spilled oil from background hydrocarbons of biogenic and pyrogenic origin (Boehm et al., 1997; Mudge, 2002), determine the weathering processes (Wang et al., 1998, 2001; Christensen et al., 2005a), and assess the ecosystem impact (risk assessment) (Porte et al., 2000; Barron and Holder, 2003).

A variety of analytical techniques have been used for oil hydrocarbon fingerprinting including gas chromatography-flame ionization detection (GC-FID), gas chromatography-mass spectrometry (GC-MS), and fluorescence spectroscopy. Oil hydrocarbon fingerprinting and spill source identification are, however, not limited to the chemical characterization using different analytical techniques, but consist of a combination of analytical techniques and methods for data preprocessing, analysis, and evaluation of the results. It is of fundamental importance for the defensibility of a combined fingerprinting approach that the process of extracting the desired information from a set of chemical data is objective and standardized. At the same time, the need to process an ever-increasing amount of samples and data implies that the data analysis should also be fast, comprehensive, and unsupervised.

Numerous methods for oil hydrocarbon fingerprinting and spill source identification have been described in the scientific literature since the 1980s (Munoz et al., 1997; Boehm et al., 1997; Burns et al., 1997; Wang et al., 1999, 2002; Stout et al., 2001; Page et al., 2002; Mudge, 2002; Daling et al., 2002; Christensen et al., 2004, 2005b, 2005d); most of them are elaborated on in this book. Standard methods are based on comparison of bulk oil properties such as total petroleum hydrocarbon concentration (TPH) (Reddy and Quinn, 1999; Wang et al., 2002), visual comparison of fluorescence spectra (Siegel et al., 1985; Siegel and Cheng, 1989) and of chromatograms obtained with GC-FID (Wang et al., 2000, 2002; Daling et al., 2002) or GC-MS (Jovancicevic et al., 1996; Ezra et al., 2000; Daling et al., 2002; Wang et al., 2002), concentrations of source-specific markers (Jovancicevic et al., 1996; Wang et al., 1999, 2002; Daling et al., 2002), bar plots of the concentrations of oil-characteristic polycyclic aromatic hydrocarbons (PAHs) (Wang et al., 1999, 2002), and

lists and double plots of diagnostic ratios of PAHs and petroleum biomarkers (Wang et al., 1999, 2002).

Most of these methods, no matter how powerful and refined, rely on skill and expertise of the analyst. Crude oils and petroleum products are complex mixtures of chemical compounds; thus, it is not feasible to identify and quantify all individual compounds in the mixture. Without using appropriate methods for data preprocessing, analyzing, and evaluating, the sheer amount of data often infers so much that few variables are chosen based on *a priori* knowledge, which in turn can largely hinder the discovery of new or more informative patterns. Likewise, some differences may be so subtle that they become invisible even to the trained eye. This can increase the level of uncertainty in conclusions of spill source identity, which then become less defensible in a court of law.

Thus, one of the most important advances in oil hydrocarbon fingerprinting is the systematic use of multivariate statistical methods for comprehensive and objective comparison and classification of oil from single and multiple sources (Christensen et al., 2004, 2005b, 2005d). Although multivariate methods have been in use for several decades in the scientific community (Jolliffe, 1986; Smilde et al., 2004), their application for oil hydrocarbon fingerprinting is relatively new and has been quite sparse. In particular, multivariate methods have been used for data analysis in organic geochemistry since the 1980s (Øygard et al., 1984; Telnaes and Dahl, 1986), but their application for oil hydrocarbon fingerprinting and spill source identification is more recent and has mostly occurred since the middle of the 1990s (Aboul-Kassim and Simoneit, 1995a, 1995b; Burns et al., 1997; Stout et al., 2001; Lavine et al., 2001; Mudge, 2002; Li et al., 2004; Christensen et al., 2004, 2005b, 2005d).

9.1.1 Multivariate Methods and Oil Fingerprinting

The first evidence of multivariate statistical methods applied to oil hydrocarbon fingerprinting dates to the beginning of the 1980s, where, for example, Øygaard et al. (1984) and Telnaes and Dahl (1986) used PCA on the normalized distribution of hopanes for oil source differentiation in geochemistry. Multivariate methods were used later for oil hydrocarbon fingerprinting in environmental forensics, and in 1995 Aboul-Kassim and Simoneit used factor analysis to study particulate fallout samples in Alexandria (Aboul-Kassim and Simoneit, 1995a) and sediment samples from the Eastern Harbor of Alexandria (Aboul-Kassim and Simoneit, 1995b). In both works, multivariate statistical tools were employed to reduce the hydrocarbon datasets into their pollution sources based on the concentrations of aliphatic and aromatic hydrocarbons. In particular, the first analysis showed that two significant factors could explain 90% of the total variation among particulate fallout samples and confirmed petrochemical (79.6%) and thermogenic/pyrolytic (10.4%) sources. Analogously, Aboul-Kassim and Simoneit (1995b) determined that untreated sewage, rather than direct inputs from boating activities or urban runoff, was the main source of petroleum hydrocarbons in the Eastern Harbor.

In 1997, Burns et al. used PCA and a least-squares iterative matching procedure to allocate PAHs in intertidal and subtidal sediment samples from the Prince William Sound of Alaska to 30 potential sources (Burns et al., 1997). PCA was used to identify 18 possible sources, including diesel oil, diesel soot, spilled crude oil in various weathering states, natural background, creosote, and combustion products from human activities and forest fires. The least-squares model was subsequently used to estimate the source mix, with the best least-squares fit of 36 PAH analytes.

In 2001, Lavine et al. employed pattern recognition and PCA to study spilled jet fuel that had undergone weathering in a subsurface environment and classified them into five types (Lavine et al., 2001). Stout et al. (2001) analyzed a suite of diagnostic PAH and biomarker ratios with PCA. The ratios were selected on the basis of high analytical precision and low susceptibility to weathering. The

analysis helped to identify the prime suspects for a heavy fuel oil (HFO) spill of unknown origin from 66 candidate sources.

In a relatively recent attempt to resolve the origin of background hydrocarbons in the sediments of Prince William Sound and the Gulf of Alaska, Mudge (2002) used partial least-squares regression. The percentage distribution of five possible sources — coal, seep oil, shales and input from two rivers — to the hydrocarbon loading in the Gulf of Alaska was estimated, with the individual contributions varying significantly across the sampling area.

Li et al. (2004) used PCA and point-to-point matching for screening and differentiation of nine oil samples based on their fluorescence emission spectra. The PCA model was able to distinguish all nine oil samples as well as the weathering extent of different oil samples, whereas only five of them were discriminated by point-to-point matching algorithms. A number of specific methods for oil hydrocarbon fingerprinting has been developed in our research group. In Christensen et al. (2005c), a semi-automatic method for processing complex first-order chromatographic data (namely GC-MS/SIM) is described that allows one to resolve convoluted peaks using mathematical functions with few parameters (i.e., Gaussian and exponential-Gaussian hybrid). The procedure was tested on chromatographic data from 20 replicate oil samples, and we found it to be less time-consuming and more objective compared to commercial software, while retaining comparable data quality. The same method was then applied in a forensic spill source identification study for the rapid and automated calculation of a large number of diagnostic ratios (Christensen et al., 2004). Four groups of diagnostic ratios derived from petroleum biomarkers (terpanes and steranes) and within homologue series of PAHs were thus evaluated simultaneously by weighted least-squares-principal component analysis (WLS-PCA), which was preferable to standard PCA as it can account for largely different variable uncertainties. The subsequent statistical testing of scores ensured an objective matching of oil spill samples with suspected source oils, and classification into positive match, probable match, and nonmatch. The sources of two spill samples (Norwegian crude oils from Oseberg East and Oseberg Field Centre) were identified and distinguished from closely related oil (crude oil from Oseberg South East).

However, some individual peaks in complex chromatograms such as the steranes at *m/z* 217 cannot be sufficiently resolved in the chromatograms to provide precise and accurate determinations of all peak areas. Furthermore, the number of possible diagnostic ratios is huge; in Christensen et al. (2004), ratios were limited to standard ratios used in geochemistry and a limited number of mostly binary ratios of alkylated naphthalenes and phenanthrenes based on pure combinatorics by calculating all possible binary isomer ratios within a compound group (e.g., 3 + 2 + 1 = 6 ratios for C1-phenanthrenes consisting of four peaks). Thus, an alternative was sought by which most of the chemical information in one or several chromatograms could be analyzed without any prior peak identification and quantification. In order to eliminate variation unrelated to chemical composition, PCA was applied to the normalized first derivatives of aligned GC-MS chromatograms of *m/z* 217 (tricyclic and tetracyclic steranes) from a selection of crude oils, petroleum products, and oil spill samples collected from the coastal environment in the weeks after the *Baltic Carrier* oil spill, March 29, 2001, Grønsund, Denmark (Christensen et al., 2005d). Four reliable components could be observed in the data, and the spill samples were correctly assigned to the corresponding sources. Due to the large number of variables, the same WLS-PCA approach previously used on diagnostic ratios was employed here. At the same time a variable selection scheme was devised to obtain the most reliable results. This method has also been tested on oil samples from an *in vitro* biodegradation experiment in which a North Sea crude oil was exposed to three mixtures of bacterial strains over a 1-year period with five sampling times (Christensen et al., 2005a). The variation in degradability within groups of isomers of methylfluorene

(*m*/*z* 180), methylphenanthrene (*m*/*z* 192), and methyldibenzothiophene (*m*/*z* 198) was used to evaluate the effects of microbial degradation on the oil composition. It was demonstrated that a mixture of strains of alkane degraders (and surfactant producers) and of PAH degraders affected the PAH isomer patterns differently than the two strain mixtures did independently.

Finally, the efficacy of an extension of PCA to higher-order data tensors for oil hydrocarbon fingerprinting was tested on fluorescence excitation-emission data (EEM) (Christensen et al., 2005b). One hundred twelve fluorescence landscapes of HFOs, light fuel oils (LFO), crude oils, lubricating oils, spill samples, and various mixed sources were analyzed simultaneously with parallel factor analysis (PARAFAC). With the exception of HFOs and crude oils, the method could discriminate between the four oil types and assign the spill samples to the corresponding sources.

9.1.2 Integrated Multivariate Oil Fingerprinting (IMOF)

A novel concept based on rapid, objective, and comprehensive multivariate statistical analysis of crude oils and refined petroleum products was developed in our research group between 2002 to 2005. The methodology has been used both for spill source identification (Christensen et al., 2004, 2005b, 2005d) and for the study of weathering of complex mixtures of oil hydrocarbons (Christensen, 2002; Christensen et al., 2005a). A number of methods are presented in this chapter as parts of the "Integrated Multivariate Oil Fingerprinting" (IMOF) methodology, which is proposed as a general framework for oil hydrocarbon fingerprinting. The IMOF methodology (Figure 9-1) is based on a combination of semiquantitative chemical analysis (cf. Section 9.2), automated and comprehensive data preprocessing (cf. Section 9.3), multivariate statistical analysis (cf. Section 9.4), and objective data evaluation (cf. Section 9.5). The data analysis step is more specifically based on multilinear decomposition methods (PCA being the most notorious, but also the parallel factor [PARAFAC] model when the data allow it) and statistical analyses, preceded by appropriate preprocessing and followed by rigorous data evaluation.

It is important to remark that IMOF is not intended as a closed set of techniques, but rather as a methodology that encompasses all aspects of oil hydrocarbon fingerprinting. Different analytical methods can provide different levels of detail of the analyzed chemical system. Hence, while in this chapter fluorescence spectroscopy is used for initial screening of oil samples and GC-MS for comprehensive compound-specific analyses, other analytical methods such as GC-FID and liquid chromatography-mass spectrometry (LC-MS) may well be incorporated into the IMOF methodology. Analogously, the goal of the chapter is not to exhaust the subject of applying multivariate statistical methods to oil fingerprinting. Nonetheless, we hope that this chapter manages to convey the great potential that these mathematical procedures have for rapid, reliable, and objective analysis of oils.

Procedures for data import, preprocessing, analysis, and evaluation in the specific methods for oil hydrocarbon fingerprinting developed in our research group and based on the IMOF methodology has been implemented in Delphi 4.0 (Borland) and Matlab 6.5 (The Mathworks). The procedure for chromatographic preprocessing of GC-MS data for analysis of complex mixtures of petroleum hydrocarbons (Christensen et al., 2005c) has been implemented in Borland Delphi 4.0 object-oriented programming, except for the extraction and sorting of data, which has been performed in Matlab 6.5 using the NetCDF software (http://my.unidata.ucar.edu). The procedures for data import, preprocessing, analysis, and evaluation in Christensen et al. (2004, 2005a, 2005b, 2005d) have been implemented in Matlab 6.5 (m-files). The relevant Matlab and Delphi files can be obtained by contacting the authors. Standard algorithms for data preprocessing and two-way data analysis can be downloaded from www.models.kvl.dk and

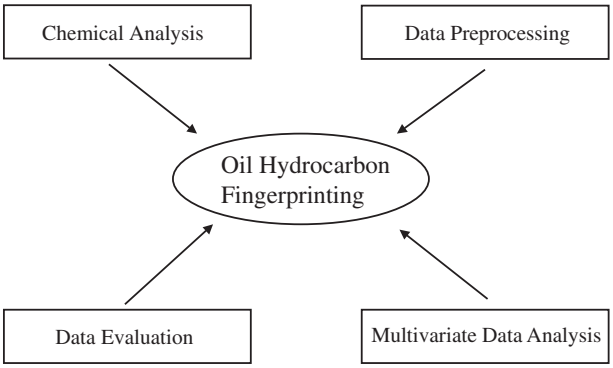

Figure 9-1 Illustration of the research strategy behind the IMOF methodology, which is based on four steps: chemical analysis, data preprocessing, multivariate data analysis and data evaluation. The aim of the work in our research groups has been to develop rapid, reliable and objective tools for comprehensive analysis and characterization of complex oil hydrocarbon mixtures using the IMOF methodology.

www-its.chem.uva.nl/research/pac. The *N*-way toolbox (Andersson and Bro, 2000) used for multiway data analyses (e.g., PARAFAC) can be downloaded from www.models.kvl.dk.

9.2 Sample Preparation and Chemical Analysis

Standard sample preparation and peak quantification procedures involve a series of time-consuming steps, such as extraction, evaporation, fractionation, addition of surrogate standards, peak detection, and integration and quantification based on response factors. The aim of this initial step of the IMOF methodology is to apply analytical procedures with limited time consumption and with consistently high-quality data. The latter can be ensured by comprehensive quality assurance and quality-control measures (QA/QC) and allows for a more adequate, comprehensive, and objective analysis of complex mixtures of oil hydrocarbons.

9.2.1 Sample Preparation

Sample preparation varies depending on the analyzed sample substrate. Thus, biota, soil, sediments, and pure oils all require different extraction, cleanup, and fractionation procedures. In our research group, the sampling steps have been limited to sampling of pure oil collected on-board ships and from vegetation and stones after oil spills (Christensen et al., 2002, 2004, 2005b, 2005d) and to sacrificing experimental units (Erlenmeyer flasks) during *in vitro* experiments (Christensen et al., 2005a). Hence, sample preparation has been limited to dilution of pure oil and extraction from stones and vegetation and to liquid/liquid extraction using dichloromethane. Subsequently, water and particles have been removed by cleanup through funnels with glass wool and sodium sulphate.

When analyzing more complex environmental matrices such as sediments, soils, and biota, semi-automated instrumental extraction techniques [e.g., microwave extraction (Jassie, 1995; Shu et al., 2000) and pressurized liquid extraction, PLE (Bandh et al., 2000)] should be preferred to less automated and more time-consuming extraction procedures such as Soxhlet extraction. PLE is especially attractive since extraction of a relatively large number of samples can be performed overnight with limited use of solvent (e.g., 10–30 ml per sample) (Bandh et al., 2000; Richter, 2000), and sample cleanup and fractionation may be integrated with the extraction procedure (Sporring et al., 2005; Nording et al., 2005). Fractionation into aliphatic, aromatic, and polar fractions, which is frequently used as a sample preparation step in oil hydrocarbon analysis (Wang et al., 1994a, 1994b), has been avoided throughout the present work to reduce the analysis time. However, chemical analysis of complex sample matrices may require this step, which can again be built into PLE procedures (Sporring et al., 2005).

9.2.2 Analytical Methods

A variety of analytical methods have been used for oil hydrocarbon analysis, including thin-layer chromatography (TLC), high-performance liquid chromatography (HPLC), gas chromatography-photo ionization detection (GC-PID), two-dimensional gas chromatography (GC-GC), gas chromatography-isotope ratio mass spectrometry (GC-IRMS), GC-FID and GC-MS; gravimetric measurements; and infrared, ultraviolet and fluorescence spectroscopy. Although all these methods have been used for oil hydrocarbon analysis, GC-FID and GC-MS are the preferred methods in most oil hydrocarbon fingerprinting studies (Aboul-Kassim and Simoneit, 1995a, 1995b; Jovancicevic et al., 1996; Munoz et al., 1997; Boehm et al., 1997; Burns et al., 1997; Wang et al., 1999, 2002; Ezra et al., 2000; Stout et al., 2001; Page et al., 2002; Mudge, 2002; Christensen et al., 2004, 2005d). Whereas GC-FID is the standard method for initial screening of oil samples, GC-MS is used for more comprehensive chemical characterization (Wang et al., 1999) since it can resolve a broad range of oil hydrocarbons including petroleum biomarkers and PAHs, and because of the low cost of quadrupole instruments. A method for initial screening of oil samples using fluorescence spectroscopy is suggested in this chapter as a complementary method to standard GC-FID screening.

9.2.3 Fluorescence Spectroscopy

Fluorescence excitation-emission matrices (EEMs) for initial screening in oil hydrocarbon fingerprinting can be obtained using a fluorescence spectrophotometer in scan mode. Fluorescence of oil is mainly caused by PAHs, which are highly fluorescent due to the presence of delocalized electrons within the aromatic rings, and because their rigid structure does not allow for efficient vibrational relaxation. Fluorescence is affected by quenching and energy-transfer processes and consequently is a complex process in multicomponent mixtures such as oil, which contains hundreds of individual PAHs.

The experimental procedure used in Christensen et al. (2005b) for fluorescence spectroscopy is based on sample dilution to avoid pronounced light absorption (i.e., the absorbances in the wavelength range between 240 and 600 nm were required to be below 0.05 absorbance units measured by UV-VIS) and thus reduces inner filter effects and effects of quenching and energy-transfer processes (Christensen et al., 2005b). The combination of high detector voltage (850 V), necessary to obtain a sufficient dilution, and a high scan speed (4800 nm/min) to reduce the analysis time, led to low signal-to-noise data. The PARAFAC model was, however, able to handle this by modeling the systematic variations and leaving the noise in the residuals.

In Christensen et al. (2005b), the EEMs were measured on a Varian Eclipse fluorescence spectrophotometer. A collection of emission scans from 250–600 nm with 2-nm increments was obtained at varying excitation wavelengths ranging from 240–475 nm with 5-nm increments. The bandwidths were 5 nm for both excitation and emission, and the scan rate was 4800 nm/min, the latter leading to a scan time of less than 10 min per sample. Each scan was comprised of 176 emission and 48 excitation wavelengths. Below an excitation wavelength of 240 nm, the spectra were noisy, and above 600 nm, the signals were negligible for the four oil types (crude oils, LFOs, HFOs, and lubricants). The fluorescence EEM measurements are thus consistent with the general goal of the IMOF methodology in being rapid and with only limited sample preparation (dilution and UV-VIS measurements). An excitation-emission scan of an HFO sample from the cargo tank of the *Baltic Carrier* is shown in Figure 9-2.

Rayleigh and Raman scatter show up in three-way fluorescence data as diagonal lines across EEMs (Andersen and Bro, 2003). The ridge due to the Rayleigh scatter is clearly visible in Figure 9-2. Rayleigh scatter is elastic (i.e., there is no energy loss); hence, the scattered emission wavelength is equal to that of excitation. Conversely, Raman scatter is

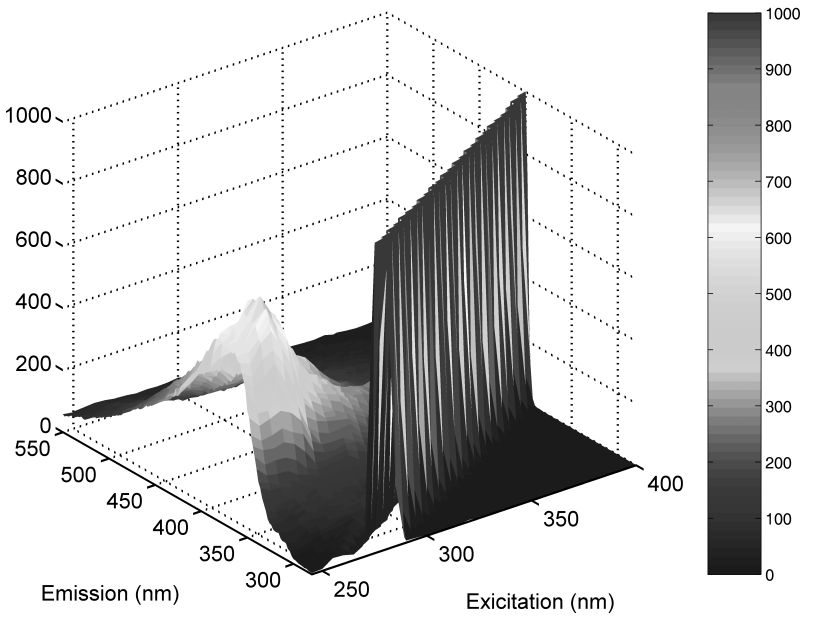

Figure 9-2 Fluorescence excitation-emission scans of an HFO. The vertical axis and scale bar show the fluorescence intensity. Modified from Christensen et al. (2005b). See color plate.

inelastic, and emission is shifted to longer wavelengths compared to excitation due to energy loss. In any case, scatter is unrelated to the chemical sample composition and cannot be modeled adequately by low-rank multilinear models (i.e., they require the fitting of numerous additional factors that have no chemical meaning) (Bro, 1997; Stedmon et al., 2003; Andersen and Bro, 2003; Christensen et al., 2005b). Several methods have been used to reduce detrimental effects due to scatter (Christensen et al., 2005b). The most immediate is to filter the corresponding signal. This can be done quite effectively by means of appropriate preprocessing before data analysis (cf. Section 9.3.3).

9.2.4 GC-MS

Mass spectra produced by GC-MS are one of the most valuable tools for identification of unknown compounds. Oil hydrocarbon fingerprinting by GC-MS is based on analysis of selected *m*/*z* ions with high sensitivity (base peaks in mass spectra) and selectivity for specific compound groups with common structure.

The selection of appropriate *m*/*z* ions for oil hydrocarbon fingerprinting in the IMOF methodology is based on prior work mainly by geochemists (Peters and Moldowan, 1993). We recommend that a large number of PAHs and petroleum biomarkers be included in the GC-MS analysis since these compounds are, generally, resistant to weathering processes and their fingerprints vary between oil sources due to the depositional environment, in-reservoir degradation, thermal maturity, and the refining process. Petroleum biomarkers and PAHs with common structures such as tri-pentacyclic terpanes, steranes, and the homologue series of C_0-C_4-phenanthrene isomers can be measured by detecting characteristic mass fragments or the molecular masses using GC-MS analysis with selected ion monitoring (SIM) (GC-MS/SIM) (Table 9-1).

A Finnigan Trace DSQ™ single quadrupole GC-MS operating in electron impact (EI) mode was used by Christensen et al. (2005d) to obtain chromatographic data for oil hydro-

Table 9-1 List of Mass Fragments (*m/z*) and Corresponding Compound Groups Analyzed with High Relevance for Oil Hydrocarbon Fingerprinting

Polycyclic Aromatic Hydrocarbons (PAHs)	*Mass Fragments* (m/z)
C_0-C_4-naphthalenes	128, 142, 156, 170, 184
C_0-C_4-phenanthrenes	178, 192, 206, 220, 234
C_0-C_3-fluorenes	166, 180, 194, 208
C_0-C_4-chrysenes	228, 242, 256, 270, 284
C_0-C_2-pyrenes and fluoranthrenes	202, 216, 230
Other PACs (5- and 6-ring)	252, 276, 278
Heterocyclic Aromatic Compounds	
C_0-C_4-benzothiophenes	134, 148, 162, 176, 190
C_0-C_4-dibenzothiophenes	184, 198, 212, 226, 240
C_0-C_1-naphthobenzothiophenes	234, 248
C_0-C_2-dibenzofuranes	168, 182, 196
Petroleum Biomarkers	
Sesquiterpanes	123
Terpanes	177, 191, 205
Steranes and diasteranes	217, 218, 259
Triaromatic steranes	231
Other Compounds	
n-alkanes and isoprenoids	85
Alkyltoluenes	105
C_0-C_3-biphenyls	154, 168, 182, 196

carbon fingerprinting. Capillary columns of 30 or 60 meters with nonpolar stationary phases (e.g., HP-5 ms, DB5, and Zebron ZB-5) are recommended for separation of oil hydrocarbons. A 60 m HP-5 ms (0.25-mm inner diameter × 0.25-µm film thickness) was used in Christensen et al. (2005d) injecting 1-µl aliquots in PTV splitless mode. The inlet temperature was 35°C during injection (1 min) and increased subsequently by 14.5°C/sec to 315°C during transfer (hold for 1 min during transfer). The column temperature program was 35°C (2 min), 60°C/min to 100°C, 5°C/min to 315°C (20 min), and transfer line and ion source temperatures: 300°C and 250°C, respectively. In Figure 9-3, a partial chromatographic profile of *m/z* 217 is shown using the injection and column temperature program just described.

Christensen et al. (2005d) suggest that increasing the sampling rate (e.g., focusing on fewer masses in each segment in the GC-MS/SIM and decreasing the dwell time) would improve the ability of the preprocessing to reduce variation unrelated to the chemical composition. For example, the initial parameterization and peak-fitting procedure applied for semiquantitative analysis would be affected positively since the number of data points in each peak would increase correspondingly (Christensen et al., 2005c). More specifically, the authors selected 48 mass fragments and analyzed them in six groups of 14–15 *m/z* ions GC-MS/SIM with a sampling rate of 1.27 scans/sec, and observed that early eluting chromatographic peaks such as naphthalene (*m/z* 128) consisted of relatively few data points (less than 10), which may be insufficient for determining the initial parameters for Gaussian peak fits (Christensen et al., 2005c). Likewise, Christensen et al. (2005d) concluded that an increased sampling rate

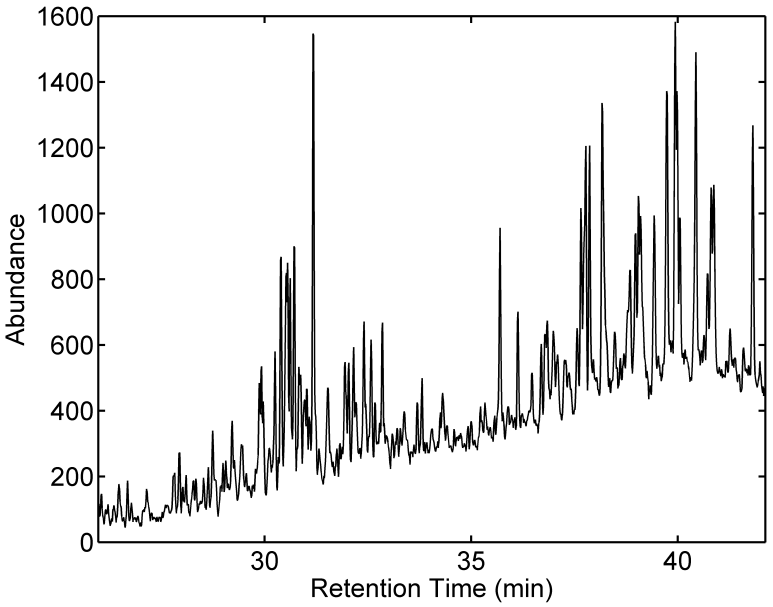

Figure 9-3 Partial GC-MS chromatogram of *m*/*z* 217, which contains tricyclic steranes (eluting between 26 and 34 min) and tetracyclic steranes (eluting between 35 and 42 min).

would allow for more refined retention time alignment. Thus, the analytical procedure was changed slightly for Christensen et al. (2005a, 2005c) such that 44 mass fragments were analyzed in 8 groups of 12 ions with a sampling rate of 2.34 scans/sec with 20 msec dwell times for each ion, which almost doubled the number of scans/sec.

9.2.5 Quality Assurance and Quality Control (QA/QC)

Semiquantitative analytical approaches are preferred to standard quantitative methods in IMOF due to the high reproducibility and low time consumption obtained with these methods. Fully quantitative approaches rely on quantification by the internal, external, or standard addition method. Conversely, a semiquantitative approach relies on frequent analyses of a laboratory reference sample, with comparable sample characteristics and chemical composition as the analytical samples. In Christensen et al. (2004, 2005a, 2005c, 2005d), the reference is a 1 : 1 mixture of a North Sea crude oil (Brent crude oil) and an HFO from the *Baltic Carrier* oil spill. In Christensen et al. (2005b), the reference is a mixture of Brent crude oil, light fuel oil (LFO), HFO, and a lubricant oil prepared in such a way that the four oils contribute approximately equally to the combined fluorescence signal. Hence, the sample characteristics of the references are comparable with those of the analytical samples (i.e., the GC-MS chromatograms contain most of the relevant peaks and the fluorescence excitation-emission spectra contain most of the fluorescence excitation-emission characteristics) of the four different oil types. The replicate analyses of laboratory reference samples have also been used for a variety of other purposes in the oil hydrocarbon fingerprinting methods elaborated in Christensen et al. (2004, 2005a, 2005c, 2005d), including calculation of the analytical uncertainty (Christensen et al., 2004, 2005a, 2005c, 2005d), optimization of data preprocessing parameters (Christensen et al., 2004, 2005a, 2005d), and normalization of diagnostic ratios (Christensen et al., 2004, 2005a, 2005c), as

well as in the peak matching procedure described in Christensen et al. (2005c).

High-quality data are a prerequisite for meaningful and reliable data analysis. This is valid for any analytical method and is particularly true for the IMOF method. Multilinear models like PCA work under a number of assumptions and requirements (for example, peaks relative to the same compound are not shifted in different measurements), and it is important that all these requirements are met to a certain extent (Tauler et al., 1995; de Juan and Tauler, 2001; Smilde et al., 2004; Christensen et al., 2004, 2005a, 2005b, 2005d). Hence, while standard quantitative methods based on peak identification and quantification are relatively unaffected by variations in peak shape (e.g., from symmetrical to tailing) and retention time shifts, PCA on sections of GC-MS/SIM chromatograms can be heavily affected by such factors due to changes in the intensity distribution of peaks (e.g., fronting, symmetrical or tailing peaks) (Christensen et al., 2005a, 2005d). Limiting these factors by comprehensive QA/QC measures are, thus, of special importance when using the latter approach. Replicated measurements of a reference oil are used, among others, for general QA/QC measures in the laboratory and thus, to some extent, to verify the compliance of the data to the requirements of the subsequent data analysis. For example, the following QA/QC measures based on the replicated references are applied in Christensen et al. (2004, 2005a, 2005c, 2005d) for QA/QC of GC-MS data:

- The chromatographic peak shapes are checked regularly at selected ion masses [e.g., *m/z* 85 (*n*-alkanes and isoprenoids), 128 (naphthalene), 180 (methylfluorenes), 191 (terpanes), 192 (methylphenanthrenes), 198 (methyldibenzothiophenes), 217 (steranes), and 252 (five-ring PACs)]. Deterioration of the chromatographic column or worsening of the conditions in the inlet (e.g., dirty liner) often causes increased tailing.
- Changes in the sensitivity of the mass spectrometer (also checked by tuning the mass spectrometer each day of analysis).
- Mass discrimination due to changes in the inlet or a dirty ion source (also checked by tuning the mass spectrometer frequently).

Changes in the chromatographic peak shapes, mass discrimination, and a significant decrease in sensitivity (e.g., more than a factor of 10) led immediately to cleaning of the ion source, change of liner and septum, or trimming of the capillary column. The septum and liner of the injector system were changed, and the ion source of the mass spectrometer cleaned regularly throughout the analytical work.

In Christensen et al. (2004, 2005a, 2005c, 2005d), the QA/QC procedure is performed once for every 50 injections (i.e., including oil samples, blanks, and references), which results in chromatographic data of consistently high quality. Thus, relative analytical standard deviations (RSD_A) based on the references were kept consistently below 3% for diagnostic ratios of well-resolved peaks (Christensen et al., 2004, 2005a, 2005c), and the variability of replicated reference samples in score plots was significantly lower than the total variability within datasets (Christensen et al., 2005d). Moreover, more than 500 injections of oils, references, blanks, and quantification standards were performed as part of the *in vitro* biodegradation study in Christensen et al. (2005a), and less than 20% of samples were reanalyzed due to insufficient data quality (i.e., poor reproducibility or sensitivity, mass discrimination, and chromatographic resolution).

9.3 Data Preprocessing

The purpose of data preprocessing in IMOF is to reduce variation in the data unrelated to the relative chemical composition such as instrumental noise, retention time shifts, analytical variability, and absolute compound concentrations. The signal-to-noise ratios are improved by reducing these factors and, consequently, facilitate the extraction of meaningful chemical information. In turn, this leads to an improved ability for the integrated oil fingerprinting method to distinguish dissimilar samples (i.e., increased resolution power) in spill source identification cases. Furthermore,

in order to be of any practical use, tools for data preprocessing have to be sufficiently rapid and should require limited human intervention. At the same time, preprocessing depends on the type of data. Thus, preprocessing of the three types of data that have been thus far analyzed in IMOF (namely, sections of GC-MS/SIM chromatograms, diagnostic ratios, and fluorescence EEM spectra) is presented separately in the ensuing sections.

9.3.1 Partial GC-MS/SIM Chromatograms

Three steps for preprocessing partial GC-MS/SIM chromatograms prior to multivariate statistical analysis were applied: baseline removal; time alignment; and normalization (Christensen et al., 2005a, 2005d). They are described in the ensuing paragraphs and have been implemented by the authors in MATLAB (The MathWorks) as one integrated methodology. The combined effects of the three preprocessing steps — baseline removal, retention time alignment, and normalization — in the oil hydrocarbon fingerprinting study (Christensen et al., 2005d) are illustrated in Figure 9-4.

9.3.1.1 Baseline Removal

Chromatographic baselines in GC-MS/SIM chromatograms are caused both by features unrelated to the chemical composition (e.g., sensitivity of the mass spectrometer) and by coelution of compounds in the complex oil mixture. Besides producing an increase in the number of factors necessary to describe the data using multilinear models, such baselines represent a problem for other preprocessing steps and should be removed prior to retention time alignment and normalization.

A number of methods can be used for this purpose. The most common are polynomial and piecewise-linear baseline fits and calculation of first or second derivatives with or without smoothing (Åberg et al., 2004). However, in complex mixtures such as crude oil and petroleum products, many peaks are not baseline-separated. Consequently, to manually select points for the baseline fit would be prone to subjective bias and is generally not viable. Although relatively refined methods exist to automate baseline removal in complex samples (e.g., using splines and convex hulls), numerical first derivatives have proven sufficient when directly applying multilinear models to oil hydrocarbon fingerprinting (Christensen et al., 2005a, 2005d). Therefore, this is the only method currently implemented in IMOF. The first derivatives (calculated through simple differences) of the partial GC-MS/SIM chromatogram (35.5–42 min) of tetracyclic steranes (*m*/*z* 217) are shown in Figure 9-4a for five reference oils and five source oils.

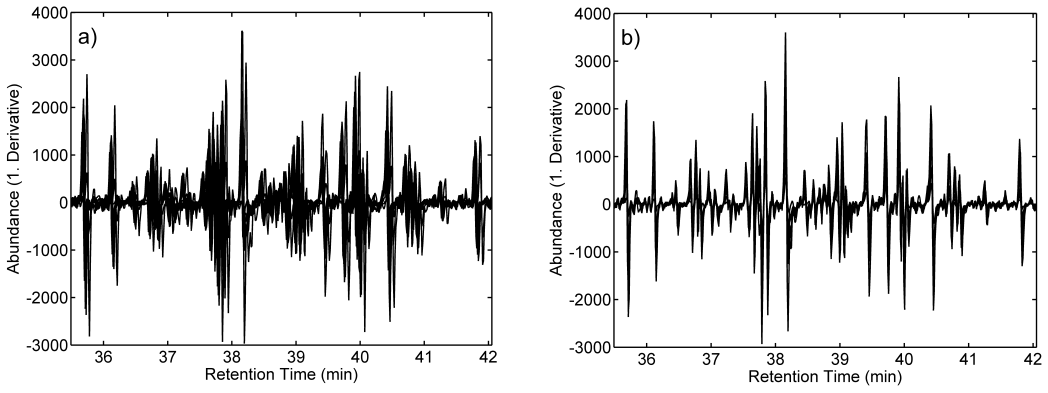

Figure 9-4 First derivative of a section of *m*/*z* 217 for five reference oils and five source oils: (a) before warping and (b) after warping using COW with segment length of 175 data points and a slack of three points. Modified from Christensen et al. (2005d).

With respect to smoothing, whether it is necessary or not when calculating the derivatives [e.g., using Savitsky-Golay (Åberg et al., 2004) or kernels] depends on the application and on the signal-to-noise ratio. Notwithstanding the fact that calculating derivatives through simple differences tends to increase noise in the data (Christensen et al., 2005d), if the signal-to-noise ratio is sufficiently good, it is still possible to extract meaningful information with limited data manipulation (Fraga et al., 2000; Johnson et al., 2003).

9.3.1.2 Retention Time Alignment

One of the most severe impediments to multivariate statistical analysis of partial GC-MS/SIM chromatograms is retention time shift, which is caused mainly by deterioration of the capillary column (Johnson et al., 2003). While recent technological advancements have helped to reduce the magnitude of this problem (Malmquist and Danielsson, 1994c; Witjes et al., 2001; Vogt and Booksh, 2004), even slight differences of one time point can affect the fitting of multilinear models (Wang and Isenhour, 1987; Malmquist and Danielsson, 1994b; Nielsen et al., 1998; Fraga et al., 2000; Rønn, 2001; Johnson et al., 2003; Eilers, 2004; Willse et al., 2005; Wong et al., 2005). Numerous methods have been proposed to correct for retention time shifts in chromatographic data in the scientific literature (van Nederkassel et al., 2005a, 2005b). While not necessarily the fastest, dynamic programming-based algorithms like dynamic time warping (DTW) and correlation optimized warping (COW) have been used with a certain degree of success for a broad range of chromatograms (Nielsen et al., 1998; Pravdova et al., 2002; Tomasi et al., 2004; van Nederkassel et al., 2005a, 2005b; Christensen et al., 2005d). While under some constraints the two algorithms may yield analogous results (Tomasi et al., 2004), only the COW algorithm is currently included in IMOF. The main reasons are that its MATLAB implementation is considerably faster than DTW and, while limited to piecewise linear corrections, COW is still capable of correcting the unwanted retention time shift for GC (Nielsen et al., 1998; Tomasi et al., 2004).

COW is a piecewise or segmented data preprocessing method, which operates on one sample at a time. It works under the assumption that corresponding subsections of two chromatograms should have the highest similarity when correctly aligned. Hence, correcting the retention time shift is reduced to identifying the boundaries of such sections and to moving them so that they occupy the same position (i.e., retention time) in all the samples as in the target.

In essence, COW aligns a sample (here a partial GC-MS/SIM chromatogram) to a target data vector (another GC-MS chromatogram) by splitting the two signals in an equal number of segments (whose length I is imposed by the analyst and depends on the data) and by finding the segments' boundaries in the sample according to a simple optimality criterion (i.e., the sum of the Pearson's correlation coefficients between corresponding segments). If the corresponding segments in the sample and in the target have different lengths, they are linearly interpolated to the same number of points. Interpolation is necessary both to be able to compute the correlation coefficient and to obtain the final alignment. Segment lengths are allowed limited changes and are restricted to integer values both to avoid extreme corrections and to keep the computational time manageable. The maximum length increase or decrease in a sample segment in terms of scan points is controlled by the slack parameter t (Eilers, 2004).

The correcting power of the algorithm is inversely proportional to the segment length and directly proportional to the slack. In order to better describe the function of COW and to illustrate when it may fail, the concept of a warping path is introduced. Alignment methods seek to find a relationship that associates the scan number (or the retention time) in the sample with a scan number in the target. This relationship, which can be an explicit function or simply represented as a set of indexes, is referred to as the *warping path* and

can be visualized in a system of axes j_{sample} versus j_{target}, where j denotes the scan number. For example, quadratic models of the type $j_{sample} = c_2 j_{target}^2 + c_1 j_{target} + c_0$, where j denotes the scan number, are sometimes enough to correct for retention time shift (Tomasi et al., 2004). On the system of axes j_{sample} versus j_{target}, this appears as a parabola. In COW the warping path is formed by segments spanning I points on the reference axis and having the slope comprised of the interval between $(I - t)I^{-1}$ and $(I + t)I^{-1}$. Thus, if the correct warping path could be approximated by a parabola, the COW warping path (given t and I) would be the optimal piecewise linear approximation of such a parabola. The shorter the segments or the larger the allowed interval for the slope, the better the approximation becomes. However, there is a lower limit for I, which must be larger than the peak width at the base to avoid artifacts and peak deformations and vice versa, large values of I can correct the retention time shift only if the curvature of the "true" warping path (i.e., the one that would yield the correct alignment) is sufficiently small and can be approximated by long segments (Christensen et al., 2005d). This has important implications when aligning GC data, because the long segments commonly used with COW (with small values of t) are likely to yield optimal results only for compounds with similar physicochemical properties that are affected in the same way by changes in the column properties (e.g., chemical changes in the stationary phase) and for m/z ratios that refer to homogeneous classes of compounds [e.g., tri- and tetracyclic steranes in m/z 217 (Christensen et al., 2005d) and methylated PAC homologues in m/z 180, 192, and 198 (Christensen et al., 2005a)]. The partial GC-MS/SIM chromatograms (m/z 217) of five reference oils and five source oils are shown in Figure 9-4, before (4a) and after (4b) alignment using COW.

Conversely, chromatographic methods with less selective detectors such as GC-FID and liquid chromatography-diode array detection (LC-DAD) produce chromatograms with more irregular retention time shifts since the chromatograms contain compounds with very different physicochemical properties. In situations with large changes in the stationary phase of the column, peaks may even change elution order, and COW cannot adequately correct for such changes. In this case, optimal alignments may infer the selection of smaller Is, mostly of the same order as the width at the base of the smallest peak one wants to align (Malmquist and Danielsson, 1994a; Andersson et al., 2004; Tomasi G. et al., 2004).

Since only integer values of the segment lengths are checked in COW, the correction is typically less fine than other methods that allow fractional corrections or express the warping path parametrically (as in the quadratic function earlier). This is visible in the residual chromatographic shift of at most one scan point after time alignment is performed (Christensen et al., 2005a, 2005d). The main consequence of these two aspects is that, while the results are deterministic, it may be necessary to test different values of I and t in order to identify the best choice. Although this is not necessarily trivial and may require supervision, a simple automated procedure based on a grid search algorithm can be used if the same sample is frequently analyzed as part of the overall procedure (cf. Section 9.2.5) (Christensen et al., 2005a, 2005d). In particular, if one disregards noise and without mean centering, the rank of a matrix consisting of perfectly aligned chromatograms obtained from the same sample is 1. In this case, the optimal values of I and t would be those that reduce to 1 the number of nonzero singular values of such a matrix. In practice, since noise is always present in experimental data, the optimal choice of warping parameters is the one that maximizes the first singular value (Christensen et al., 2005a, 2005d).

Note that the residual shift mentioned earlier has an important secondary effect. Namely, as the rank of the fitted model (i.e., the number of principal components for a matrix) increases, the additional factors explain less and less of the chemical variation and more of the relatively small variation caused by residual shifts (Johansson et al., 1984).

9.3.1.3 Normalization

The chromatographic abundances are related to the sensitivity of the instrument and concentrations of the single constituents. It is affected by several parameters such as sampling, extraction, cleanup, fractionation, and the sensitivity of the mass spectrometer. Variations in data due to these factors are likely to mask the compositional information in the subsequent data analyses. Hence, normalization is a prerequisite for objective and automated comparison of sections of preprocessed GC-MS/SIM chromatograms by multivariate statistical analyses. Normalization to a constant Euclidean norm is a common procedure used to compensate for concentration effects and sensitivity changes [Eq. (9-1)]:

$$x_{nj}^{N} = \frac{x_{nj}}{\sqrt{\sum_{j=1}^{J} x_{nj}^{2}}} \tag{9-1}$$

where x_{nj} is the first derivative of the nth chromatogram at the jth retention time and J is the total number of retention times. This method is affected by the so-called closure of the dataset (i.e., if one peak increases, the size of the other peaks decrease) (Christensen et al., 2005d). In situations where the amount of information is limited (i.e., few peaks), this may lead to correlations in data that are only present due to the closure, which makes chemical interpretation of the results more difficult. In such cases, more complex normalization schemes using only a limited set of retention times referring to specific peaks could be adopted. To reduce the uncertainty and closure effects, these peaks should preferably be large and relatively constant in oil samples of different origin. Under such circumstances, the denominator in Eq. (9-1) is modified to the sum of the squared first derivatives of selected retention times instead of all retention times.

9.3.2 Diagnostic Ratios

As opposed to sections of GC-MS/SIM chromatograms, diagnostic ratios condense the compositional information in a smaller number of variables. Three types of diagnostic ratios are suggested within the IMOF methodology. They are listed in Eqs. (9-2)–(9-4).

$$DR = a_n^S / a_{n^*}^S \tag{9-2}$$

$$DR = a_n^S / \left(a_n^S + a_{n^*}^S\right) \tag{9-3}$$

$$DR = \frac{DR^S}{\left(DR^S + DR^R\right)}$$
$$DR^S = a_n^S / a_{n^*}^S \quad \text{or} \quad DR^S = a_n^S / \left(a_n^S + a_{n^*}^S\right)$$
$$DR^R = a_n^R / a_{n^*}^R \quad \text{or} \quad DR^R = a_n^R / \left(a_n^R + a_{n^*}^R\right) \tag{9-4}$$

where a_n^s and $a_{n^*}^s$ are the peak areas, peak heights, or concentrations of compound n and n*, respectively, or the sums of several compounds in an oil sample, and a_n^R and $a_{n^*}^R$ are the peak areas, peak heights, or concentrations of the corresponding compound(s) in a reference oil sample. If the peak values denote compound concentrations, the method of determining the diagnostic ratios is named "quantitative"; conversely, if peak areas or heights are employed, the procedure is termed "semiquantitative." The use of semiquantitative analyses as opposed to quantitative ones has been a matter of debate in the scientific community. The quantitative approach is often preferred to the semiquantitative one for oil hydrocarbon fingerprinting purposes (Wang et al., 1999). The main reasons are the inherent normalization based on the internal quantification procedure, which is generally thought to be more precise than a semiquantitative approach, and the potential for unraveling mixed oil signatures. In our experience, this is not necessarily the case, especially when employing multivariate statistical tools to comprehensive collections of diagnostic ratios or partial GC-MS/SIM chromatograms as described in Section 9.3.1. Furthermore, the semiquantitative method has proven to be sufficiently precise, less time-consuming, and simpler to implement (Christensen et al., 2004). In a recent oil hydrocarbon fingerprinting study, 137 diagnostic ratios based on petroleum biomarkers and PAHs were calculated

from semiquantitative data (peak areas) (Christensen et al., 2004). RSD_A varied between 0.09% and 5.1% using Eq. (9-3) and decreased to 0.05% to 3.2% with external normalization to the oil reference analyzed closest in time [Eq. (9-4)]. In Christensen et al. (2005c), 72 diagnostic ratios were calculated from semiquantitative peak data (peak areas and heights). The RSD_A were below 5% for all diagnostic ratios based on well-separated and baseline distorted peaks using a new method based on peak modeling using mathematical formulas as well as commercial quantification software for peak quantification. For 10 diagnostic ratios based on incompletely resolved peaks, the RSD_A were between 2% and 8%. Finally, in an *in vitro* biodegradation study (Christensen et al., 2005a), the RSD_A for 19 diagnostic PAH ratios calculated from Eq. (9-3) varied between 0.5% and 4.8%. Direct comparisons between quantitative and semiquantitative approaches for calculating diagnostic ratios are difficult, since the RSD_As are rarely listed in oil hydrocarbon studies. However, in a study by Stout et al. (2001), the RSD_As of petroleum biomarker ratios were generally comparable or higher than those obtained in our studies. Note, however, that no matter which approach is used to calculate diagnostic ratios, the RSD_A should always be calculated as part of the QA/QC procedure and for justifying the results on sample identity.

9.3.3 Preprocessing of Fluorescence Spectra

Fluorescence data require an entirely different pretreatment than chromatographic ones. However, the purpose is still to reduce information that is not related to the chemical composition (i.e., what is relevant to oil hydrocarbon fingerprinting). The main problem with EEM signals is the presence of Rayleigh and Raman scatter in the data (e.g., Figure 9-2). These two physical phenomena are particularly problematic when modeling the data using low-rank multilinear models (i.e., PARAFAC), and how to remove their influence has been the subject of recent research (Andersen and Bro, 2003; Rinnan, 2004).

There are several ways of dealing with Rayleigh scatter; the most common is to insert missing values in a more or less wide diagonal band in the EEM landscape centered on the wavelengths at which such scatter is observed (i.e., at emission wavelengths close to the wavelength of the exciting light — see Figures 9-2 and 9-5) (Rinnan, 2004; Christensen et al., 2005b). Most algorithms for fitting multilinear models are capable of handling missing values using an expectation maximization approach (Bro, 1998; Tomasi and Bro, 2006). Conversely, Raman scatter effects are reduced, but not eliminated, by subtracting blanks (e.g., recorded in dichloromethane) (Stedmon et al., 2003; Tomasi and Bro, 2005).

As shown in Section 9.4, the PARAFAC decomposition essentially requires that the underlying model that gives rise to the data be inherently multilinear. Among others, this means that, in order for the mathematical model to be appropriate, the emission spectrum for one compound (or class of substances) must remain unaltered at all excitation wavelengths and in all samples. Likewise, that excitation spectrum must be the same at all emission wavelengths and in all samples (Andersen and Bro, 2003). While indeed an approximation, severe deviations from these assumptions may have important consequences on the usefulness of the fitted models (Bro, 1998). For this reason, instrument biases in EEM data are reduced by applying an excitation/emission correction spectrum derived from a combination of a Rhodamine spectrum and the spectrum from a ground quartz diffuser (Christensen et al., 2005b). The excitation/emission correction removes small artifacts in EEMs due to variations in detector efficiency as a function of wavelength. A preprocessed EEM is shown in Figure 9-5.

The fact that the signal recorded at emission wavelengths lower than the excitation one is physically zero may also negatively interact with the way EEM signals are modeled by PARAFAC (Andersen and Bro, 2003). Thus,

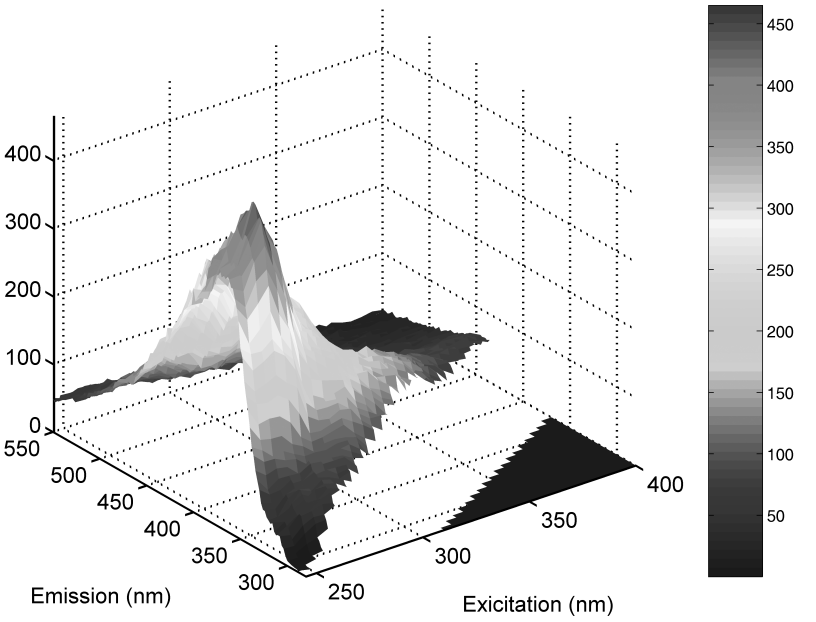

Figure 9-5 Preprocessed fluorescence EEM of the same oil sample as shown in Figure 9-2 after blank subtraction, insertion of missing values, and a small triangle of zeros and excitation/emission correction. The vertical axis and scale bar show the fluorescence intensity. Modified from Christensen et al. (2005b). See color plate.

the fact that the observed value is zero implies that for all compounds in a sample, the emission factor or the extinction coefficient at the corresponding wavelengths must be zero (assuming that negative contributions would make no sense from a physical point of view). It is obvious that this assumption is untenable for most fluorescent compounds, and some preprocessing is required. As for scatter, weighting or setting these values to missing are the most obvious choices. While it may be tempting to set all values "below" the Rayleigh scatter ridge to missing, this has proved to be problematic as the different factors are somehow allowed to interact in the missing values area, giving rise to artifacts and lack of convergence (Rinnan, 2004; Thygesen et al., 2004; Tomasi and Bro, 2005). The zeros can be considered to have a stabilizing effect, and the width of the missing values band should be the subject of optimization (Thygesen et al., 2004).

Finally, note that constraining the model parameters to be nonnegative or with certain choices of the loss function (i.e., instead of setting specific array elements to missing, assign a nonzero small weight to them) may also be beneficial (Bro et al., 2002; Andersen and Bro, 2003; Rinnan, 2004). Whether these choices are preferable to preprocessing depends on the data. We refer the reader to the original literature for further readings on the subject.

9.4 Multivariate Statistical Data Analysis

Technological advancement has made it possible to produce and treat datasets of increasing size and complexity. However, the number of underlying factors that give rise to a certain dataset is typically limited (within a relatively homogeneous dataset) compared to the number of observed variables. For example, a univariate chromatogram (GC-MS/SIM) may be considered as a vector of data in which each element corresponds to a certain scan time and may well contain thousands of data points, but

the number of factors actually present in such data is at most equal to the number of peaks in the chromatogram. This is even more evident for a multivariate signal (as is the case for GC-MS/Full Scan data), where several tens or hundreds of measurements (m/z values) are taken at each time point in the chromatogram or when several samples are stacked into a matrix (or for EEM fluorescence data). To some extent, the peak quantification and calculation of diagnostic ratios, which often precedes the data analysis in oil hydrocarbon fingerprinting, serve the purpose of reducing the number of variables. The concept can be further extended if one takes into account that even the relative concentrations of the compounds may be related to one another in a way that may result from some common physicochemical phenomenon (e.g., depositional environment, thermal maturity, and in-reservoir degradation) affecting the composition. For example, the heptadecane/pristane and octadecane/phytane diagnostic ratios involve compounds whose physicochemical properties are almost identical and both describe biodegradation (because the branched alkanes at the denominator are less susceptible to microbial degradation than the *n*-alkanes that are at the numerator).

Multivariate statistical methods such as PCA and PARAFAC can be used to determine the most salient sources of variation and condense redundant information related to collinear variables. In this fashion, more complex patterns that might have been unknown or somewhat hidden prior to the data analysis may emerge and lead to a better understanding of the dataset and of the underlying factors. Moreover, these models have the capability of filtering some of the noise always present in experimental data, which may confound the data interpretation.

An important aspect of multilinear modeling is the determination of the correct number of factors (the so-called rank of the model). A number of tools exist for this purpose; some of them are generally applicable independently of the model (like cross-validation or the presence of relevant systematic variation in the residuals) while others may rely on the specific properties of a given model (like core consistency or the split-half analysis for PARAFAC) (Jolliffe, 1986; Bro, 1998; Rinnan, 2004).

It is also worth mentioning that the sources of systematic variation unrelated to the chemical composition should be reduced to a minimum prior to the multivariate analyses as suggested in the IMOF methodology. Otherwise, an increased number of factors are obtained whose interpretation will be difficult from a chemical point of view. Likewise, it may be necessary to select a subset of the variables or to downscale those with the highest level of noise in order to let certain phenomena become more pronounced. The main problem could be that the amount of information in the chosen dataset is insufficient to clearly establish important patterns in data. Additional information, such as the analytical and sampling uncertainties, can be used for automatic and objective selection or scaling of variables (cf. Section 9.4.2). This point is particularly relevant for complex mixtures like crude oil and petroleum products that have been exposed to weathering processes.

A necessary condition that is often overlooked is that, in order to yield interpretable results, the model that one decides to use has to be appropriate for the data at hand. For example, nonlinear behaviors cannot be modeled adequately using few principal components, and PARAFAC does not yield physically interpretable results if there are significant interactions between the factors. If these conditions are not met, it may be necessary to use a different model (Smilde et al., 2004). The appropriateness of the used model can be diagnosed by looking at the magnitude and systematicity of residuals which contain the variation not described by the model or at the scores and loadings. In general, if a multivariate model does not describe data sufficiently well or fails to identify systematic chemical information, the interpretation can lead to inadequate and even misleading conclusions in, for example, environmental forensic investigations. In this sense, in order to soundly establish the correctness of one's intuition, it may be necessary to validate the results

based on *a priori* knowledge as well as testing the results from a statistical point of view (cf. Section 9.5.2). In the next two sections, PCA and the PARAFAC decomposition are outlined, as they have been used in IMOF.

9.4.1 Multilinear Models

9.4.1.1 Two-Way Case

A broad range of bilinear models exists for matrices (i.e., two-way datasets comprised of samples × variables) for both decomposition and calibration methods. The most common in chemometrics are PCA, principal component regression (PCR), partial least-squares regression (PLSR), and multivariate curve resolution (MCR) (Jolliffe, 1986; Wold et al., 1987; Martens and Næs, 1996; Smilde et al., 2004). In bilinear models, the basic assumption is that the systematic variation in data matrix **X** of dimensions $n \times p$ can be described by the product of two matrices of size $n \times F$ and $p \times F$ (Martens and Næs, 1996; Smilde et al., 2004). The different models then vary in how these two matrices are determined. In PCA, **X** is decomposed in the product of a score matrix **T** with a loading matrix **P** (with elements t_{if} and p_{jf}) that are column-wise orthogonal plus a matrix of residuals, **E** with elements e_{ij} (Bro, 1997) [see Eq. (9-5)]:

$$x_{ij} = \sum_{f=1}^{F} t_{if} p_{jf} + e_{ij} \quad (i = 1 \ldots I; j = 1 \ldots J)$$
$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{9-5}$$

where x_{ij} is the data point for the i-th sample and j-th variable, F is the number of principal components (PC's) and T identifies transposition. F is also referred to as the mathematical rank of the model, and the minimum value of F for which the residuals are zero is the mathematical rank of the **X**. Moreover, the PCs are ordered in **T** and **P** in decreasing order of variance, and the columns of **P** have a norm equal to 1. Hence, the first PC is the most relevant source of variation, the second PC is the second-most relevant source, and so forth, and for any given value of F the sum of the square of the residuals is minimized. All these conditions on the scores and loading matrices imply that the solution is uniquely determined. However, for F larger than 1, there is an infinite number of solutions that describe the data equally well (i.e., with the same value of the sum of squared residuals) that can be obtained from appropriate linear combinations of the columns of **T** and **P** (Bro, 1997; Riu and Bro, 2003). This is commonly referred to as "rotational freedom" and implies that there is no guarantee that the PCs also resemble the original factors that underlie the data. A clarifying example is provided by spectral data. If there are F spectrally active chemical species in a sufficiently diluted solution and the absorption at p wavelengths is measured for n samples, the physical model for the observed data is of the type described by Eq. (9-5), where t_{if} would be proportional to the concentration of the fth species in the ith sample, and p_{jf} would be proportional to the extinction coefficient at the jth recorded wavelength. Under these conditions, each column of the matrix **P** contains the absorption spectrum of one of the chemical species present in the solution and the columns of **T** of the corresponding concentrations. However, these spectra cannot be orthogonal, because orthogonality also implies that a varying number of elements in **P** are negative. In fact, MCR methods seek a rotation of the PCs such that all the columns of **P** are nonnegative.[1]

The PCs are themselves linear combinations of the original variables and can be obtained through a variety of methods. The most reliable from a numerical point of view is the singular value decomposition (SVD). The main problem with this method is that in its standard form it cannot handle missing values, which are a rather common occurrence. When these are present, other algorithms based on, for example, expectation maximization or weighted least squares, should be used (Grung

[1] Note that even this condition is generally not sufficient for uniqueness (Smilde et al., 2004).

and Manne, 1998; Walczak and Massart, 2001).

A final note regards the application of PCA (and other bilinear models) to higher-order tensors, which requires that the elements of the data array be rearranged in a matrix. However, while this so-called matricization is always possible, it also leads to a dramatic increase of the number of parameters compared to models such as PARAFAC that exploit the additional structural relationships that may be present in the dataset. In particular, for an $n \times p \times q$ array matriced to an $n \times pq$ matrix, each PC entails the estimation of $n + pq$ parameters as opposed to, for example, PARAFAC, which requires only $n + p + q$.

9.4.1.2 Higher-Order Arrays

As mentioned in the previous paragraph, bilinear models may require the estimation of an exceeding number of parameters (and likely incur overfitting) when applied to higher-order arrays (tensors) whose elements have been rearranged in a matrix. There are several extensions to the multiway case of PCA and of the other bilinear models (e.g., PARAFAC, Tucker, and *n*PLS). However, not all the properties of bilinear models are retained when they are extended to higher orders, and new ones may emerge. Hence, a bit of care is required to determine which is the best choice for a given problem. For example, Tucker and PARAFAC models can both be considered extensions of PCA, but have considerably different properties (Smilde et al., 2004). Due to space restrictions, only PARAFAC will be described in more detail. We refer the reader to the cited literature for additional information about Tucker and *n*PLS models and their relation with PARAFAC.

PARAFAC is perhaps the most straightforward extension of bilinear models and the PARAFAC equation for each element of the data array is obtained by simply adding a term for each additional order after the second. Thus, for a three-way array $\underline{\mathbf{X}}$ of dimensions $I \times J \times K$ (e.g., obtained by stacking on top of each other $J \times K$ EEM fluorescence measurements relative to I samples), the equation for PARAFAC is:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \quad (i = 1 \ldots I; j = 1 \ldots J; k = 1 \ldots K)$$
$$\underline{\mathbf{X}} = \mathbf{A} \circ \mathbf{B} \circ \mathbf{C} + \underline{\mathbf{E}} \tag{9-6}$$

where x_{ijk} and e_{ijk}, respectively, identify array elements and the corresponding residuals, and $\circ$ symbolizes the tensor product (Smilde et al., 2004). In order to highlight the symmetry in the role of these three matrices, they are oftentimes referred to solely as loading matrices (Bro, 1997; Thygesen et al., 2004). The appellation of scores is sometimes reserved to the mode relative to the sample being analyzed.

Unlike PCA, **A**, **B**, and **C** need not be columnwise orthogonal for the model to be identifiable under rather mild conditions (e.g., that no two columns of any of the loading matrices be proportional) up to trivial factor permutations and scaling (Smilde et al., 2004; Thygesen et al., 2004). The fact that orthogonality is not necessary to uniquely determine the model parameters has important consequences. The most important is that, provided that a minimum in the loss function is found and the rank of the model is correct, the model parameters can have, for several chemical data, a straightforward physical interpretation (Leurgans and Ross, 1992). For EEM fluorescence measurements of sufficiently diluted solutions, and if I is number of samples, J the number of emission wavelengths, and K the number of excitation wavelengths, the elements of the fth column of the first loading matrix are proportional to the concentration of one of the fluorescent species in the solution and the fth columns of **B** and **C** contain the emission and excitation spectrum, respectively, of such fluorophores (Leurgans and Ross, 1992; Andersen and Bro, 2003).

This identification between the PARAFAC model and the physical model, however, is subject to a number of restrictions, the first of which is that there are no observable interactions between the chemical species that are

modeled by separate factors, and that the profile relative to such species in one mode does not change as the variable in the other modes changes. Thus, for EEM data, it is necessary that little or no quenching occurs and that emission and excitation spectra for the same fluorophore are identical in all samples. However, experimental data often deviate from these assumptions, which leads to inaccurate estimates for the physical model parameters. In some cases, it may be useful to enforce some constraints when they are known to apply (e.g., nonnegativity and unimodality on one or more of the loading vectors) (Bro, 1997; Stedmon et al., 2003; Andersen and Bro, 2003).

It is worth mentioning that, while constraints may improve the robustness of the modeling and are in accordance with chemical *a priori* knowledge (i.e., negative concentrations and fluorescence intensities have no physical meaning), they are by no means necessary to yield an interpretable solution provided that the deviations from the basic multilinearity assumptions are not too large. On the contrary, such artifacts may be quite informative on the presence of outliers and on the effect of the pattern of missing values. A detailed introduction to PARAFAC and examples of its applications to analysis of fluorescence EEMs have recently been published in relation to oil hydrocarbon fingerprinting (Christensen et al., 2005b).

Finally, as of this publication, there is no algorithm of fixed complexity that can be used to fit PARAFAC to three-way data that contain noise and, for particularly difficult problems, it may not be possible to retrieve a meaningful solution (Tomasi and Bro, 2006). One of the problems that adds to the difficulty of fitting a PARAFAC model is that PARAFAC models of increasing rank are not nested. That is, the first component in a rank-2 PARAFAC model is not equal to the only component of a rank-1 PARAFAC model. Hence, it is not possible, as it is in PCA, to fit a model with a large number of factors and then select the correct rank. Determination of the appropriate number of factors has mainly been based on split-half analysis, analysis of residuals (Bro, 1997, 1998; Andersen and Bro, 2003; Christensen et al., 2005b), and comparison of PARAFAC factors with EEMs of individual PAHs (Christensen et al., 2005b).

9.4.2 Variable Selection and Scaling

Although data are preprocessed prior to the multivariate statistical data analysis (Figures 9-4 and 9-5), some variables (e.g., chromatographic intensities, diagnostic ratios, or excitation–emission pairs) are more informative than others. There are at least two possible solutions to this problem. One is to select the variables that should be retained in the data array and exclude the noisiest, less informative ones; a second one is to downscale these variables or single measurements using a weighted least-squares fitting criterion (WLS). Both approaches have been tested as parts of IMOF (Christensen et al., 2004, 2005d).

The resolution power (r), intended as the ability of distinguishing dissimilar samples while keeping identical ones as close as possible to one another in the space defined by the principal components (score plots), has been used in IMOF as a criterion for optimizing variable selection and scaling (Christensen et al., 2005a, 2005d). Ideally, an optimal resolution power is obtained for the combination of variables with lowest variation over replicate samples compared to the total variance explained by the model (i.e., minimal r).

One option used in IMOF is to define r as the ratio between the variance of the PCA scores of replicate samples (or of samples that belong to a certain class) and the overall variance explained by the same model and to progressively include additional variables according to their analytical uncertainty (Christensen et al., 2005d). This method was applied to a section of m/z 217 chromatograms (i.e., associated to tri- and tetracyclic sterane biomarkers) and allowed to reduce the number of variables from 1251 to 351 with a twofold improvement on the resolution power for samples that had not been included in the calibration set. A somewhat more statistical treat-

ment of the subject can be found in Pierce et al. (2005), in which ANOVA is used for feature selection.

Another option for selecting diagnostic ratios for oil hydrocarbon fingerprinting was suggested in Christensen et al. (2004) using the concept of diagnostic power (DP). The selection of diagnostic PAH and biomarker ratios for oil hydrocarbon fingerprinting was based on the analytical precision of individual ratios as well as their resistances to weathering (Christensen et al., 2004). The DP can be calculated from Eq. (9-7), and ratios with lowest DP can be excluded from the data analysis to reduce the analytical noise and sampling variability that can have negative effects on the spill source identification.

$$DP = \frac{RSD_V}{RSD_A} \quad \text{or} \quad DP = \frac{RSD_V}{RSD_S} \qquad (9\text{-}7)$$

where the relative sampling standard deviation (RSD_S) is the combined random errors from the chemical analysis as well as the sample variation (relative sampling standard deviation). The latter includes the sample heterogeneity and the effect of weathering. DP is defined as the relative standard deviation of a diagnostic ratio in oils with different origin (RSD_V) divided by RSD_A or RSD_S. The variable selection method is based on the diagnostic powers used in Christensen et al. (2004) to select the most descriptive biomarker ratios for oil hydrocarbon fingerprinting. RSD_S, RSD_V, and DP were calculated for 17 diagnostic ratios and sorted with respect to their DP (Table 9-2). A comparable approach was suggested by Stout et al. (2001), who excluded ratios with RSD_A larger than 5% from the PCA.

Variable selection can be a rather time-consuming process that may introduce some degree of subjectivity in the data analysis. Therefore, a second method of improving the performance of PCA has been investigated in which all the variables are retained but are inversely scaled according to their analytical

Table 9-2 Application of a Variable-Outlier Detection Method to 17 Biomarker Ratios

Diagnostic Ratios	RSD_S	RSD_V	*DP*
C24TT/ (C24TT + 30ab)	1.5	37.7	25.7
25nor30ab / (25nor30ab + 29ab + 30ab)	2.2	43.3	20.0
25nor30ab / (25nor30ab + Ts+Tm)	1.9	33.5	17.8
C24TT / (C24TT + 29ab + 30ab)	2.2	34.6	15.9
(29ab+30ab)/(29ab+30ab + 27bbSt(S+R) + 28bbSt(S+R))	0.4	4.8	11.8
30G / (30G+30ab)	1.7	19.3	11.6
Ts / (Ts + Tm)	1.2	9.7	10.5
RC27TA / (RC27TA+ RC28TA)	1.0	8.3	8.5
32abS / (32abS + 32abR)	0.3	2.7	8.0
29ab / (29ab + 30ab)	1.5	10.0	6.9
29bb(R+S) / (29bb(R+S) + 29aa(R+S))	1.0	6.9	6.9
27db(R+S) / (27db(R+S) + 27bb(R+S))	2.7	15.3	5.7
29aaS / (29aaS + 29aaR)	1.7	8.4	5.0
28ab / (28ab+30ab)	6.7	29.0	4.3
29bbSt(S+R) / (27bbSt(S+R) + 28bbSt(S+R) + 29bbSt(S+R))	1.6	6.3	3.9
27bbSt(S+R) / (27bbSt(S+R) + 28bbSt(S+R) + 29bbSt(S+R))	1.5	3.5	2.4
28bbSt(S+R) / (27bbSt(S+R) + 28bbSt(S+R) + 29bbSt(S+R))	1.8	4.0	2.2

The values have been estimated from 24 oil spill samples from the *Baltic Carrier* oil spill (Christensen et al., 2004). The DPs were calculated from RSD_S and RSD_V using Eq. (9-7), whereas the diagnostic ratios were double normalized to the reference oil analyzed closest in time using Eq. (9-4). The normalization factor in Eq. (9-4) is omitted in the description for brevity. The individual compound abbreviations have been defined in the supporting information of Christensen et al. (2004).

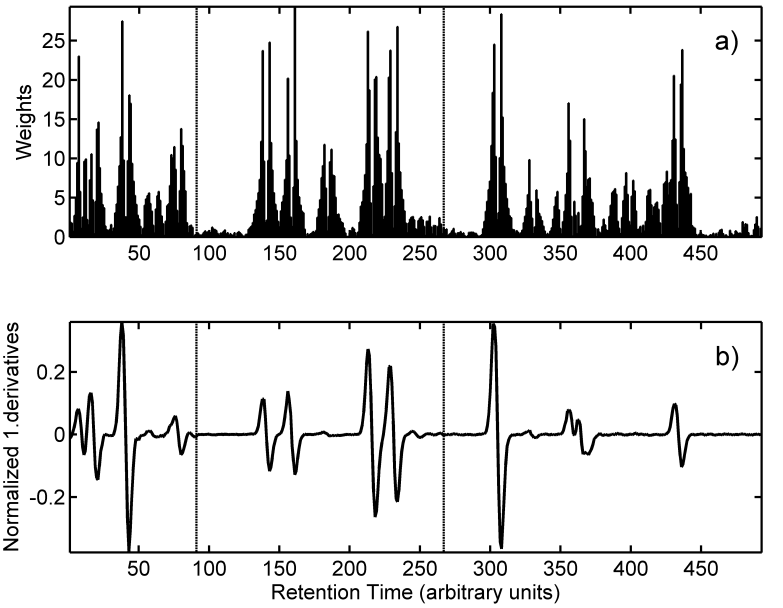

Figure 9-6 (a) Weights (RSD_A^{-1}) used for WLS-PCA and (b) the aligned and normalized mean combined chromatogram of unweathered Brent crude oil. The mean was calculated from four replicate samples. Modified from Christensen et al. (2005a).

(or sampling) uncertainty (Christensen et al., 2005a, 2005d). For example, three partial GC-MS/SIM chromatograms, C_1-fluorenes (*m/z* 180), C_1-phenanthrenes (*m/z* 192), and C_1-dibenzothiophenes (*m/z* 198), were combined and analyzed in Christensen et al. (2005a) to evaluate the effects of microbial degradation of oil in an *in vitro* experiment. The inverses of RSD_As (weights) were calculated from the aligned and normalized reference set consisting of 33 replicate references, using the optimal warping parameters and the complete chromatographic sections for normalization (90, 175, and 225 data points). The weights are shown in Figure 9-6a and were used in the WLS-PCA to scale the importance of chromatographic intensities with respect to their analytical uncertainty. The first derivative of the aligned and normalized mean combined chromatogram of the unweathered Brent crude oil, used in Christensen et al. (2005a), are, for comparison, shown in Figure 9-6b. The peak and noise regions have, respectively, high and low weights. Thus, the importances of peak regions are high (but not the peak maxima, which are characterized by a higher uncertainty), compared to those of noise regions during fitting of the PCA model.

In Christensen et al. (2004, 2005a, 2005d), the WLS-PCA model gave comparable resolution power as the optimal variable selection followed by PCA. Hence, WLS-PCA was found to be a more attractive method, since it is highly objective and no data are excluded from the analysis. Yet it is important to acknowledge the fact that the weights should describe intrinsic properties of the dataset (e.g., analytical uncertainty). If this is not the case, WLS-PCA will result in a model with lower resolution power than PCA without variable selection and even to bias in the data. Algorithms for weighted PCA can be downloaded from www.models.kvl.dk and http://www-its.chem.uva.nl/research/.

9.5 Data Evaluation

A large number of methods can be used to evaluate the outputs from multivariate statistical analyses and to classify and match sus-

pected source oils and spill samples in oil hydrocarbon fingerprinting. The methods described in the following sections are all based on the scores and loadings from multilinear models (again PCA and PARAFAC).

9.5.1 Visual Inspection of Score and Loading Plots

The most straightforward and intuitive method for evaluating a multilinear model is through score and loading plots, which may reveal relations between multiple samples and variables, respectively. The importance of visual inspection and comparison of score and loading plots can be illustrated from the correlations observed between scores and loadings in a recent oil hydrocarbon fingerprinting study, where preprocessed sections of GC-MS/SIM chromatograms of *m*/*z* 217 (i.e., tri- and tetracyclic steranes) for 101 oil samples were analyzed by WLS-PCA (Christensen et al., 2005d). The sample set used in this study was divided into a calibration set of 61 chromatograms from 51 source oils and 10 replicate samples (61 × 1231), a reference set containing 18 replicate reference samples (18 × 1231), and a test set comprised of 16 weathered oil samples collected in the spill area after the *Baltic Carrier* oil spill (Christensen et al., 2005d) and two spilled oils (analyzed in triplicate) from the study of Faksness et al. (2002). WLS-PCA was applied to the mean-centered calibration set, and the number of significant principal components was found by visual inspection of the chromatographic loadings, and confirmed by evaluating the variability of replicate samples (Christensen et al., 2005d). The visual inspection of loadings showed that some peaks in PC5 and subsequent principal components described the residual misalignment rather than systematic changes in chemical composition. Residual shifts show up in the cumulative sum of loadings as first derivative peaks. As an increasing number of principal components is extracted from data, the ratio between the systematic information and the variations caused by insufficient alignment increases, until the latter becomes the most pronounced and the subsequent components describe residual shifts and additional instrumental noise.

It can be concluded from visual inspection and evaluation of the PC1 to PC4 scores in this study (e.g., Figure 9-7) that the *Baltic Carrier* oil spill samples and the corresponding source oil were clustered in PC1 through 4, despite weathering for up to 14 days. Likewise, two Round-Robin spill samples (Faksness et al., 2002), Spill I and Spill II, were grouped along the principal components with the corresponding sources, Oseberg East (E), and Oseberg Field Centre (FC). It is important to remark that, while principal components are ordered according to their explained variance and the differences along the first components are in general more significant than along subsequent ones, later components are not necessarily of lesser importance for the separation of dissimilar oil samples. Especially for large datasets that may require a larger number of components to be properly described, an important separation between similar samples is often found in later components. Note, for example, that although oil samples from Oseberg South East (SE) and Oseberg FC are closely related, and the samples could not be separated along the three most significant components (i.e., PC1 to PC3) (Christensen et al., 2005d), they were easily separated along PC4 (see Figure 9-7), the latter describing only 6.27% of the total variation in the dataset.

Interpretation of the correlations between samples in score plots, by visual inspection of the loading plots, further facilitates comparisons of source oils and spill samples. Loading plots show which original variables (e.g., retention times and diagnostic ratios) are responsible for the directions, changes, and groupings observed in the corresponding score plots, and their importance for oil hydrocarbon fingerprinting can be illustrated from Figure 9-8, which shows the cumulative sum of the fourth principal component, found to distinguish two closely related North Sea crude oils.

The PC4 loadings in Figure 9-8 are negative for ββ-isomers of C_{27} to C_{29}-rearranged steranes and positive for αα-isomers. The ratio of

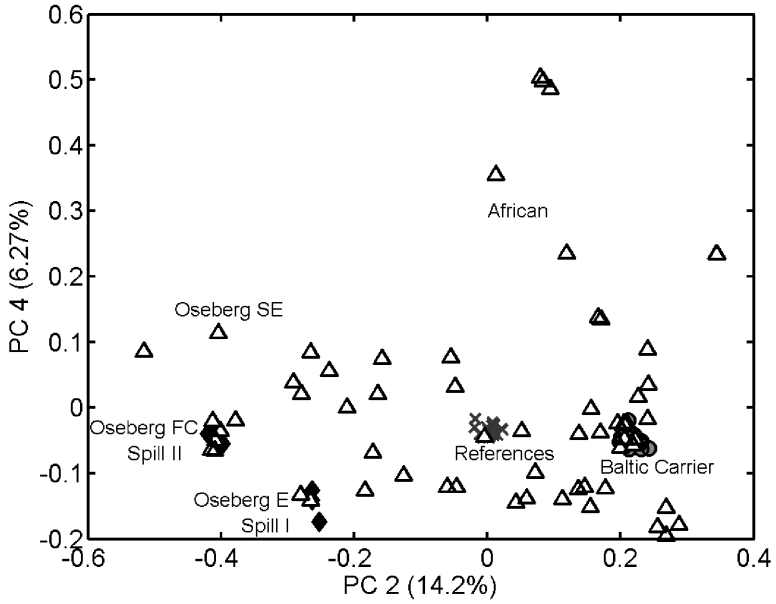

Figure 9-7 Score plot (PV2 versus PC4) using WLS-PCA on sections of preprocessed partial GC-MS/SIM chromatograms of *m/z* 217. The PCA model was calculated from the calibration set (61 × 1231), whereas a reference set (18 × 1231) of replicate references and test set (22 × 1231) were calculated by projecting the data onto the loadings. The test set was comprised of 16 *Baltic Carrier* oil spill samples and two spill samples from a Round-Robin exercise analyzed in triplicate (Spill I and Spill II) (Faksness et al., 2002).

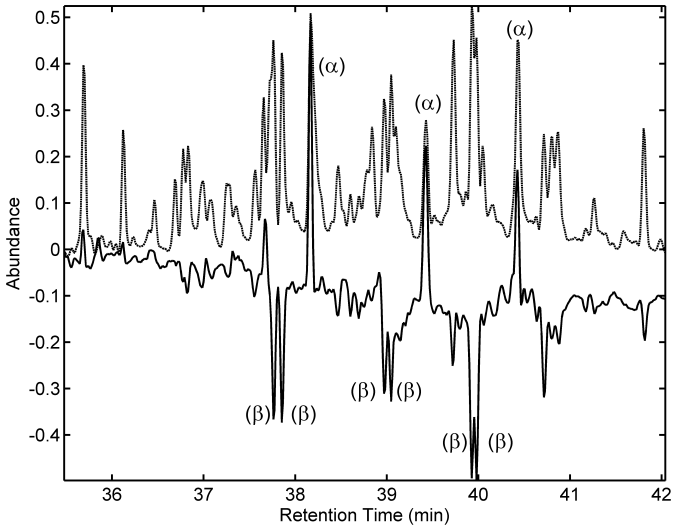

Figure 9-8 Integrated mean-chromatogram (dotted line) and integrated PC4 loadings (solid line) for WLS-PCA in the oil hydrocarbon fingerprinting study in Christensen et al. (2005d). The αα-steranes (αα) and ββ-steranes (ββ) are marked in the plot.

C_{29}-rearranged steranes (ββ / (ββ + αα)) is a highly specific parameter for maturity and appears to be independent of source organic matter input (Peters and Moldowan, 1993). The ββ isomers have a higher thermal stability than αα isomers; thus, the above ratio increases with thermal maturity. Since, the PC4 loadings are negative for ββ isomers and positive for αα isomers, the ββ / (ββ + αα) ratios are highest in oils located at low PC4 in

the score plot (Figure 9-7). Thus, the main separation of oils from Oseberg SE and Oseberg FC can be explained by a higher thermal maturity of the oil from Oseberg FC. Likewise, oils from Oseberg E are slightly more mature than those from Oseberg FC, and the African crude oils studied have positive PC4 scores and thus have low relative maturity with respect to the oils in the calibration set.

The loadings of PC1, PC2, and PC3 (not shown) can be used to facilitate the correlation and differentiation of source oils and spill samples along PC1 to 3 in the same way as PC4 was used. The three components describe boiling point range, clay content of source rock, and carbon number distribution of sterols in the organic matter of the source rock, respectively (Christensen et al., 2005d). Thus, all four significant principal compounds in this study could be interpreted directly from *a priori* knowledge of the effects of source rock depositional environment and thermal maturity and the refining process on the sterane composition seen as variations in the chromatographic profile of *m/z* 217 (corresponding to the composition of tri- and tetracyclic steranes) (Christensen et al., 2005d).

When closely related source oils are present in a large dataset, it is likely that major trends, which are represented by the first PCs, mask the differences between these oils. The components that describe these minor differences may not be included in the optimal PCA model, as they represent a minimal variation as compared to the total (Jolliffe, 1986). To ensure that important but small variations are not masked by major trends, PCA can be applied to a subset of source oils that lie close to the spilled oil along the components retained in the original PCA model. In Christensen et al. (2004), local PCA models (i.e., modeling a subset of closely located samples) have been used to focus the data analysis on separating related samples with similar biomarker and PAH composition. Hence, the first few components describe variations relevant for the specific oil spill case by separating related samples in the first few PCs instead of in higher PCs. The latter are more affected by noise since they describe only a small percentage of the total variation in the dataset.

In Christensen et al. (2005b), normalized PARAFAC scores are used to characterize and match oil samples based on their relative PAH composition. The four oil types (crude oils, HFOs, LFOs, and lubricants) could, except for some overlap of HFOs and crude oils, be distinguished from two score plots where Factor 3 versus Factor 5 is shown in Figure 9-9. The evaluation showed that crude oils have the most regular distribution of factors and a large variability in the content of high-molecular-weight PAHs. In contrast, LFOs and lubricants have a high relative content of low-molecular-weight PAHs, whereas HFOs have a high relative content of high-molecular-weight compounds.

Since PARAFAC factors are essentially unique up to trivial permutation and scaling, if, upon validation [e.g., through split-half analysis and cross-validation (Smilde et al., 2004)], they appear stable and reliable, they can be expected to reflect some underlying phenomena that give rise to the data. In the case of EEM fluorescence, these phenomena can most often be assimilated to emission and excitation spectra of homogeneous classes of compounds (or even single fluorophores). In Christensen et al. (2005b), the PARAFAC factors were interpreted by comparing excitation and emission loadings with EEMs of selected PAHs and the fluorescence characteristics for a broad range of PAHs (Figure 9-10). The graphical comparison of PARAFAC loadings with the spectra of specific fluorescent compounds was sufficient to support the original hypothesis on the meaning of the factors and allowed to associate them to mixtures of PAHs with similar fluorescence characteristics: a mixture of naphthalenes and dibenzothiophenes, fluorenes, phenanthrenes, chrysenes, and five-ring PAHs.

9.5.2 Numerical Comparisons and Statistical Tests

The objectivity of the matching process of spill and source oil samples can be improved by

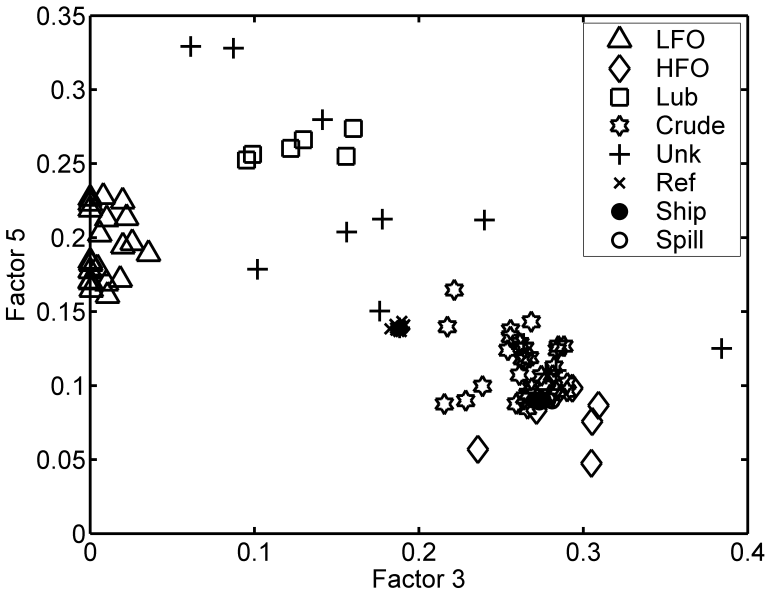

Figure 9-9 PARAFAC score plot. Factor 3 versus Factor 5. Symbols for LFOs, HFOs, lubes, crude oils, unknown oil samples, replicate reference oils, and the triplicate spill and ship samples are explained in the legend. The 17 replicate references are circled, and arrows mark the position of spill and *Baltic Carrier* ship samples.

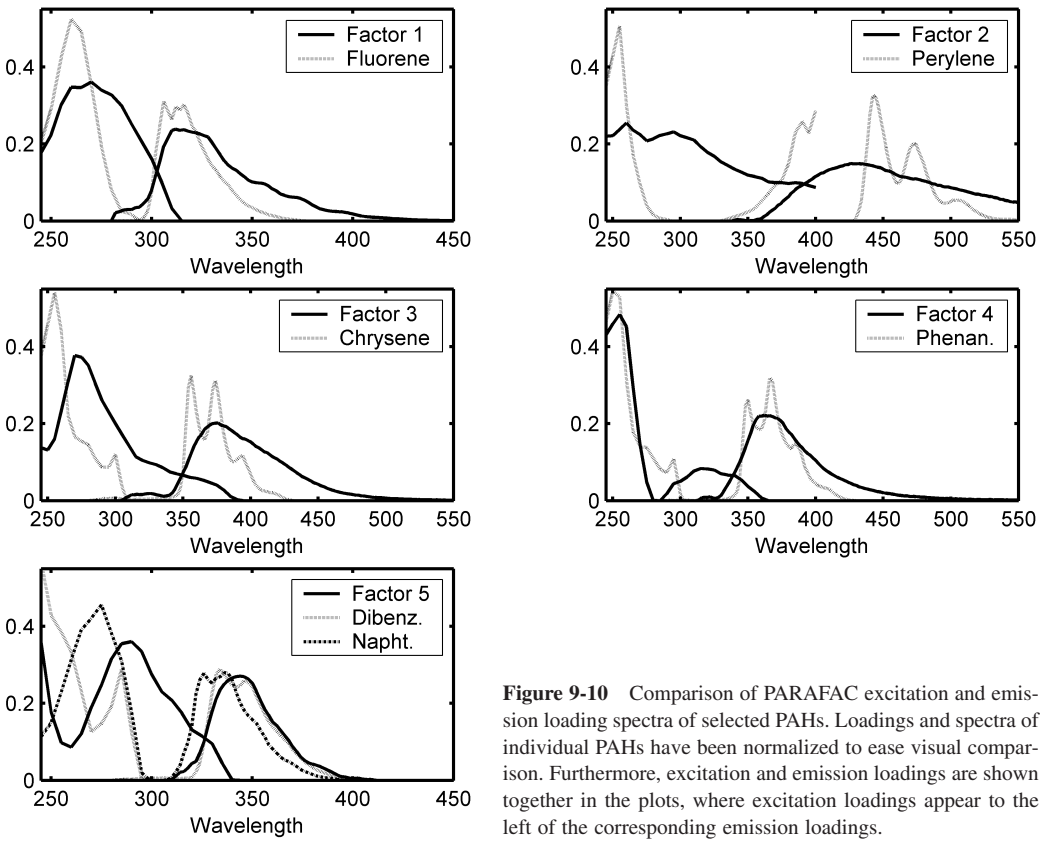

Figure 9-10 Comparison of PARAFAC excitation and emission loading spectra of selected PAHs. Loadings and spectra of individual PAHs have been normalized to ease visual comparison. Furthermore, excitation and emission loadings are shown together in the plots, where excitation loadings appear to the left of the corresponding emission loadings.

numerical comparisons (e.g., correlation coefficient) or statistical tests. Using the former approach, the similarity of oil samples can be calculated from the scores (Christensen et al., 2005b) or by point-to-point matching (Li et al., 2004).

In Christensen et al. (2005b), the similarity of oil samples is calculated using the correlation coefficient based on the similarity of normalized scores. More specifically, an oil sample collected in the spill area two weeks after the *Baltic Carrier* spill accident was compared to oil samples in the database. It was found that the triplicate samples from the cargo tank of the *Baltic Carrier* and three additional HFOs gave the highest match to spill samples ($r = 0.998 - 0.999$). Comparisons based on PCA scores could also have been used in Christensen et al. (2005) for objective spill/source matching of preprocessed chromatographic sections of biomarker hydrocarbons.

An even more objective method for matching oil samples is applied in Christensen et al. (2004) and is as such a highly objective alternative to visual inspection of score and loading plots as well as the use of similarity indices. The method consists of statistical evaluation based on the overall null hypothesis (H_0) that the spilled oil and the tested source oil are identical. The optimal number of principal components in a multivariate model (i.e., the retained components or factors) can be tested independently accepting a certain error level (often 5%, $\alpha = 0.05$). The method is used in Christensen et al. (2004) by independently testing the significant principal components using the inequality in Eq. (9-8) and accepting an error level of 5%, $\alpha = 0.05$. If the inequality is false in at least one of these tests, the overall H_0 is rejected and the tested source oil is "beyond reasonable doubt" not the source of the spill.

$$\frac{\left|\bar{t}_k^{(\text{spill})} - \bar{t}_k^{(\text{source})}\right|}{s_k^{(\text{pooled})}\sqrt{\frac{1}{n_{\text{spill}}} + \frac{1}{n_{\text{source}}}}} \leq q_{\alpha,\text{d.f.}} \qquad (9\text{-}8)$$

where $s_k^{(\text{pooled})}$ is the pooled standard deviation, which can be calculated from either the analytical or the sampling standard deviation. In addition, n_{spill} and n_{source} are the number of replicates used to calculate the mean scores along the k-th principal component of the spilled oil ($\bar{t}_k^{(\text{spill})}$) and a source oil ($\bar{t}_k^{(\text{source})}$). $q_{\alpha,d.f.}$ is the α-quantile from t-student's distribution with $d.f.$ degrees of freedom and $s_k^{(\text{pooled})}$ the pooled standard deviation for the k-th principal component.

There can be several possible outcomes of the classification of source oil with respect to the spilled oil in the multiple tests. Here are the criteria used in Christensen et al. (2004). However, other criteria can be used, such as those suggested in the modified Nordtest methodology (Daling et al., 2002).

Positive match: H_0 is acceptable (5% error level) for the tested source oil and the spill sample, and H_0 is rejected for all other source oils in the dataset.

Probable match: H_0 are acceptable for the tested source oil and spill sample, but the same holds for other source oils.

Nonmatch: H_0 is rejected.

The match criteria just described are based entirely on statistics and require a consistently high data quality. It is important to emphasize that caution against an overreliance on such unsupervised classification based on purely statistical criteria needs to be taken, and any matches need to be evaluated in light of other available data (e.g., additional compound groups).

9.6 Conclusions and Perspectives

Rapid, reliable, and objective tools are a requirement for the characterization of complex chemical mixtures such as oil. This chapter describes the development of such tools for oil hydrocarbon fingerprinting and spill source identification. An integrated multivariate oil hydrocarbon fingerprinting (IMOF) methodology comprised of four steps is described throughout the chapter: sample preparation and chemical analysis, data preprocessing, multivariate statistical analysis, and data evaluation. Figure 9-11 gives a schematic presentation of the specific oil hydrocarbon fingerprinting methods devel-

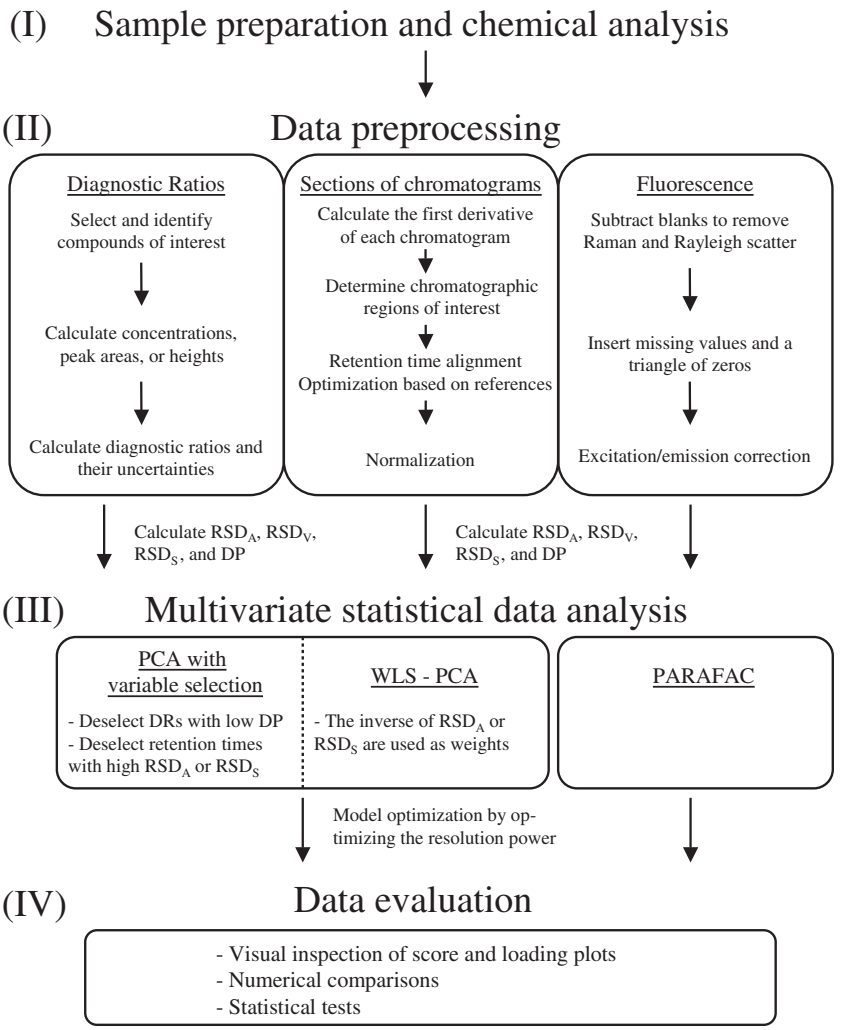

Figure 9-11 Flowchart for the IMOF methodology including the individual oil hydrocarbon fingerprinting methods developed by Christensen et al. (2004, 2005b, 2005c, 2005d).

oped in our research group and based on the IMOF framework.

The chemical analyses were based on fluorescence spectroscopy and GC-MS/SIM. The fluorescence spectroscopy method facilitated a rapid (less than 10 min per sample) screening and characterization of oil samples based on their main composition of PAHs. The technique provides a rapid and alternative, yet complementary (revealing PAH groups), technique to the more traditional GC-FID screening (revealing mainly paraffins). GC-MS/SIM has been used for more comprehensive compound-specific analysis. The GC-MS fingerprinting technique was based on a semiquantitative approach including frequent analysis of a reference sample. The semiquantitative approach, used to calculate diagnostic ratios, was rapid and with similar precision as a fully quantitative approach based on extensive use of internal and quantification standards. The replicate analysis of reference samples could furthermore be used for QA/QC, uncertainty estimations, automating the preprocessing, and external normalization. The semiquantitative approach based on GC-

MS/SIM constituted an important part of the IMOF methodology.

The preprocessing tools comprised semi-automated peak identification and quantification (Christensen et al., 2004, 2005c), analysis of sections of chromatograms by baseline removal, time warping and normalization (Christensen et al., 2005a, 2005d), and automated preprocessing of fluorescence EEMs (Christensen et al., 2005b). Time warping combined with PCA is a rapid and objective approach for oil hydrocarbon analysis compared to peak identification and quantification. The results of the research in our laboratory show, however, that the time warping approach is more affected by changes in the data quality than is peak quantification. Hence, the use of time warping in routine investigations requires extensive QA/QC measures to be taken. Furthermore, the ratio between the inherent variability in the dataset and the variability due to insufficient alignment are important criteria to determine whether or not the method provides an appropriate preprocessing tool for oil analysis.

The use of multivariate statistical methods such as PCA and PARAFAC are the cornerstone of the IMOF methodology. The multivariate methods enable the analysis and assessment of large datasets by extracting a number of principal components or factors that describe the prominent trends in data. A refined and more objective data analysis was obtained by WLS-PCA compared to PCA with variable selection. Yet, the weights should describe intrinsic properties of the dataset such as the analytical uncertainty. If this is not the case, the subjectivity of the data analysis increases, and the model becomes increasingly biased.

Rapid and objective oil hydrocarbon fingerprinting was attained using several evaluation techniques. Although excellent for monitoring and assessing the fate of complex oil hydrocarbon mixtures, visual interpretation of score and loading plots are often insufficient for proper analysis of chemical fingerprinting data. More objective methods for defensibly linking spilled oil with possible sources in an oil database are also presented in this chapter. The analytical and sampling uncertainties have been used in Christensen et al. (2004) to test the null hypothesis and determine the source of spill samples. Thus, the conclusions are less dependent on data variations and subjective decisions. In summary, each of the four steps in the IMOF methodology contributes to rapid, objective, and comprehensive analyses of complex oil hydrocarbon mixtures.

The methods developed in our research group and described in this chapter can be employed in routine investigations, and they appear as significant improvements compared to standard qualitative and quantitative methods for oil hydrocarbon fingerprinting. The limited human intervention required — and the extended amounts of chemical information that can be generated, analyzed, and evaluated — are the major and obvious strengths of the IMOF methodology. More specifically, the methods described in Christensen et al. (2004, 2005b, 2005d) enable a more comprehensive and objective matching of oil samples than the standard methods, especially if the spilled oils have chemical characteristics related to several suspected source candidates. Furthermore, a sample database can be built over time that allows each new spill to be compared to the ever-growing database without the need for reanalysis or reprocessing of existing data. These methods can easily be implemented and used for routine investigations in forensic oil spill laboratories. Fluorescence spectroscopy combined with PARAFAC can be used for prescreening oil samples (Christensen et al., 2005b), while GC-MS/SIM combined with fast and objective preprocessing, data analysis, and data evaluation (Christensen et al., 2004, 2005d) can be used for compound-specific fingerprinting.

Acknowledgments

The authors acknowledge the co-authors of a recent oil hydrocarbon fingerprinting article as well as Lotte Frederiksen, Jørgen Avnskjold, and Peter Christensen for technical assistance. The work with developing the IMOF technology was financed by Roskilde University, the

National Environmental Research Institute, the Natural Sciences Research Foundation, all from Denmark, and the European Commission (contracts "BIOSTIMUL," QLRT-1999-00326, and "ALARM," GOCE-CT-2003–506675).

References

Åberg, K.M., R.J.O. Torgrip, and S.P. Jacobsson, Extensions to peak alignment using reduced set mapping. Classification of LC/UV data from peptide mapping. *J. Chemometrics*, 2004, **18**, 465–473.

Aboul-Kassim, T.A.T. and B.R.T. Simoneit, Aliphatic and aromatic-hydrocarbons in particulate fallout of Alexandria, Egypt — sources and implications. *Environmental Science & Tech.*, 1995a, **29**(10), 2473–2483.

Aboul-Kassim, T.A.T. and B.R.T. Simoneit, Petroleum hydrocarbon fingerprinting and sediment transport assessed by molecular biomarker and multivariate statistical-analyses in the Eastern harbor of Alexandria, Egypt. *Marine Pollution Bull.*, 1995b, **30**(1), 63–73.

Andersen, C.M. and R. Bro, Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *J. Chemometrics*, 2003, **17**(4), 200–215.

Andersson, C.A. and R. Bro, The N-way toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, 2000, **52**(1), 1–4.

Andersson, F.O., R. Kaiser, and S.P. Jacobsson, Data preprocessing by wavelets and genetic algorithms for enhanced multivariate analysis of LC peptide mapping. *J. Pharmaceutical and Biomedical Analysis*, 2004, **34**(3), 531–541.

Bandh, C., E. Bjorklund, L. Mathiasson, C. Naf, and Y. Zebuhr, Comparison of accelerated solvent extraction and soxhlet extraction for the determination of PCBs in Baltic Sea sediments. *Environmental Science & Tech.*, 2000, **34**(23), 4995–5000.

Barron, M.G. and E. Holder, Are exposure and ecological risks of PAHs underestimated at petroleum contaminated sites? *Human and Ecological Risk Assessment*, 2003, **9**(6), 1533–1545.

Boehm, P.D., G.S. Douglas, W.A. Burns, P.J. Mankiewicz, D.S. Page, and A.E. Bence, Application of petroleum hydrocarbon chemical fingerprinting and allocation techniques after the *Exxon Valdez* oil spill. *Marine Pollution Bull.*, 1997, **34**(8), 599–613.

Bro, R., PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 1997, **38**(2), 149–171.

Bro, R., Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications, Ph.D., University of Amsterdam, 1998.

Bro, R., N.D. Sidiropoulos, and A.K. Smilde, Maximum likelihood fitting using ordinary least squares algorithms. *J. Chemometrics*, 2002, **16**(8–10), 387–400.

Burns, W.A., P.J. Mankiewicz, A.E. Bence, D.S. Page, and K.R. Parker, A principal-component and least-squares method for allocating polycyclic aromatic hydrocarbons in sediment to multiple sources. *Environ. Toxicology and Chem.*, 1997, **16**(6), 1119–1131.

Christensen, J.H., Application of multivariate data analysis for assessing the early fate of petrogenic compounds in the marine environment following the *Baltic Carrier* oil spill. *Polycyclic Aromatic Compounds*, 2002, **22**(3–4), 703–714.

Christensen, J.H., A.B. Hansen, G. Tomasi, J. Mortensen, and O. Andersen, Integrated methodology for forensic oil spill identification. *Environmental Science & Tech.*, 2004, **38**(10), 2912–2918.

Christensen, J.H., A.B. Hansen, U. Karlson, J. Mortensen, and O. Andersen, Multivariate statistical methods for evaluating biodegradation of mineral oil. *J. Chromatography A*, 2005a, **1090**(1–2), 133–145.

Christensen, J.H., A.B. Hansen, J. Mortensen, and O. Andersen, Characterization and matching of oil samples using fluorescence spectroscopy and parallel factor analysis. *Analytical Chem.*, 2005b, **77**(7), 2210–2217.

Christensen, J.H., J. Mortensen, A.B. Hansen, and O. Andersen, Chromatographic preprocessing of GC-MS data for analysis of complex chemical mixtures. *J. Chromatography A*, 2005c, **1062**(1), 113–123.

Christensen, J.H., G. Tomasi, and A.B. Hansen, Chemical fingerprinting of petroleum biomarkers using time warping and PCA. *Environ. Sci. Tech.*, 2005d, **39**(1), 255–260.

Daling, P.S., L.G. Faksness, A.B. Hansen, and S.A. Stout, Improved and standardized methodology for oil spill fingerprinting. *Environ. Forensics*, 2002, **3**(3–4), 263–278.

de Juan, A. and R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets. *J. Chemometrics*, 2001, **15**(10), 749–772.

Eilers, P.H.C., Parametric time warping. *Analytical Chem.*, 2004, **76**, 404–411.

Ezra, S., S. Feinstein, I. Pelly, D. Bauman, and I. Miloslavsky, Weathering of fuel oil spill on the east Mediterranean coast, Ashdod, Israel. *Organic Geochem.*, 2000, **31**(12), 1733–1741.

Faksness, L.G., P.S. Daling, and A.B. Hansen, Round Robin study — Oil spill identification. *Environ. Forensics*, 2002, **3**(3–4), 279–291.

Fraga, C.G., B.J. Prazen, and R.E. Synovec, Comprehensive two-dimensional gas chromatography and chemometrics for the high-speed quantitative analysis of aromatic isomers in a jet fuel using the standard addition method and an objective retention time alignment algorithm. *Anal. Chem.*, 2000, **72**(17), 4154–4162.

Grung, B. and R. Manne, Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1998, **42**, 125–139.

Jassie, L., Microwave technology in the analysis of contamination by petroleum. *Intl. Laboratory News*, 1995, 18.

Johansson, E., S. Wold, and K. Sjodin, Minimizing effects of closure on analytical data. *Anal. Chem.*, 1984, **56**(9), 1685–1688.

Johnson, K.J., B.W. Wright, K.H. Jarman, and R.E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J. Chromatography A*, 2003, **996**(1–2), 141–155.

Jolliffe, I.T., *Principal Component Analysis*, New York: Springer-Verlag, 1986.

Jovancicevic, B.S., L.Z. Tasic, P.S. Polic, J.M. Nedeljkovic, A.K. Golovko, and D.K. Vitorovic, GC-MS in crude oil correlation studies — effects of biodegradation on sterane and terpane maturation parameters. *J. Serbian Chem. Soc.*, 1996, **61**(9), 817–821.

Lavine, B.K., D. Brzozowski, A.J. Moores, C.E. Davidson, and H.T. Mayfield, Genetic algorithm for fuel spill identification. *Analytica Chimica Acta*, 2001, **437**, 233–246.

Leurgans, S. and R.T. Ross, Multilinear models: Applications in spectroscopy. *Statistical Sci.*, 1992, **7**(3), 289–319.

Li, J.F., S. Fuller, J. Cattle, C.P. Way, and D.B. Hibbert, Matching fluorescence spectra of oil spills with spectra from suspect sources. *Analytica Chimica Acta*, 2004, **514**(1), 51–56.

Malmquist, G. and R. Danielsson, Alignment of chromatographic profiles for principal component analysis — a prerequisite for fingerprinting methods. *J. Chromatography A*, 1994, **687**(1), 71–88.

Martens, H. and T. Næs, *Multivariate Calibration*. Chichester, UK: John Wiley & Sons, 1996.

Mudge, S.M., Reassessment of the hydrocarbons in Prince William Sound and the Gulf of Alaska: Identifying the source using partial least-squares. *Environ. Sci. & Tech.*, 2002, **36**(11), 2354–2360.

Munoz, D., P. Doumenq, M. Guiliano, F. Jacquot, P. Scherrer, and G. Mille, New approach to study of spilled crude oils using high resolution GC-MS (SIM) and metastable reaction monitoring GC-MS-MS. *Talanta*, 1997, **45**(1), 1–12.

Nielsen, N.P.V., J.M. Carstensen, and J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatography A*, 1998, **805**(1–2), 17–35.

Nording, M., S. Sporring, K. Wiberg, E. Bjorklund, and P. Haglund, Monitoring dioxins in food and feedstuffs using accelerated solvent extraction with a novel integrated carbon fractionation cell in combination with a CAFLUX bioassay. *Anal. Bioanal. Chem.*, 2005, **381**(7), 1472–1475.

Øygard, K., O. Grahl-Nielsen, and S. Ulvøen, Oil/oil correlation by aid of chemometrics. *Organic Geochem.*, 1984, **6**, 561–567.

Page, D.S., A.E. Bence, W.A. Burns, P.D. Boehm, J.S. Brown, and G.S. Douglas, A holistic approach to hydrocarbon source allocation in the subtidal sediments of Prince William Sound, Alaska, embayments. *Environ. Forensics*, 2002, **3**(3–4), 331–340.

Peters, K.E. and J.M. Moldowan, *The Biomarker Guide: Interpreting Molecular Fossils in Petroleum and Ancient Sediments*. Englewood Cliffs, NJ: Prentice Hall, 1993.

Pierce, K.M., J.L. Hope, K.J. Johnson, B.W. Wright, and R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *J. Chromatography A*, 2005, **1096**(1–2), 101–110.

Porte, C., X. Biosca, M. Sole, and J. Albaiges, The Aegean Sea oil spill on the Galician Coast (NW Spain). III: The assessment of long-term sublethal effects on mussels. *Biomarkers*, 2000, **5**(6), 436–446.

Pravdova, V., B. Walczak, and D.L. Massart, A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, 2002, **456**(1), 77–92.

Reddy, C.M. and J.G. Quinn, GC-MS analysis of total petroleum hydrocarbons and polycyclic aromatic hydrocarbons in seawater samples after the North Cape oil spill. *Marine Pollution Bull.*, 1999, **38**(2), 126–135.

Richter, B.E., Extraction of hydrocarbon contamination from soils using accelerated solvent extraction. *J. Chromatography A*, 2000, **874**(2), 217–224.

Rinnan, Å., Application of PARAFAC on spectral data, Ph.D., The Royal Veterinary and Agricultural University, 2004.

Riu, J. and R. Bro, Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems*, 2003, **65**(1), 35–49.

Rønn, B.B., Nonparametric maximum likelihood estimation for shifted curves. *J. Royal Stat. Soc., Series B* (*Statistical Methodology*), 2001, **63**(2), 243–259.

Shu, Y.Y., R.C. Lao, C.H. Chiu, and R. Turle, Analysis of polycyclic aromatic hydrocarbons in sediment reference materials by microwave-assisted extraction. *Chemosphere*, 2000, **41**(11), 1709–1716.

Siegel, J.A. and N.Z. Cheng, Fluorescence of petroleum-products 4. Three-dimensional fluorescence plots and capillary gas-chromatography of midrange petroleum-products. *J. Forensic Sciences*, 1989, **34**(5), 1128–1155.

Siegel, J.A., J. Fisher, C. Gilna, A. Spadafora, and D. Krupp, Fluorescence of petroleum products 1. Three-dimensional fluorescence plots of motor oils and lubricants. *J. Forensic Sciences*, 1985, **30**(3), 741–759.

Smilde, A.K., R. Bro, and P. Geladi, *Multi-Way Analysis. Applications in the Chemical Sciences*. Chichester, England: John Wiley & Sons Ltd, 2004.

Sporring, S., S. Bowadt, B. Svensmark, and E. Bjorklund, Comprehensive comparison of classic Soxhlet extraction with Soxtec extraction, ultrasonication extraction, supercritical fluid extraction, microwave assisted extraction and accelerated solvent extraction for the determination of polychlorinated biphenyls in soil. *J. Chromatography A*, 2005, **1090**(1–2), 1–9.

Stedmon, C.A., S. Markager, and R. Bro, Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Marine Chem.*, 2003, **82**(3–4), 239–254.

Stout, S.A., A.D. Uhler, and K.J. McCarthy, A strategy and methodology for defensibly correlating spilled oil to source candidates. *Environ. Forensics*, 2001, **2**(1), 87–98.

Tauler, R., A. Smilde, and B. Kowalski, Selectivity, local rank, 3-way data analysis and ambiguity in multivariate curve resolution. *J. Chemometrics*, 1995, **9**(1), 31–58.

Telnaes, N. and B. Dahl, Oil-oil correlation using multivariate techniques. *Organic Geochem.*, 1986, **10**(1–3), 425–432.

Thygesen, L.G., A. Rinnan, S. Barsberg, and J.K.S. Moller, Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area. *Chemometrics and Intelligent Laboratory Systems*, 2004, **71**(2), 97–106.

Tomasi, G. and R. Bro, A comparison of algorithms for fitting the PARAFAC model. *Computational Stat. Data Anal.*, 2006, **50**(7), 1700–1734.

Tomasi, G., F. van den Berg, and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometr.*, 2004, **18**, 231–241.

Tomasi, G. and R. Bro, PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems*, 2005, **75**(2), 163–180.

van Nederkassel, A.M., M. Daszykowski, D.L. Massart, and Y. Vander Heyden, Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling. *J. Chromatography A*, 2005a, **1096**(1–2), 176–186.

van Nederkassel, A.M., V. Vijverman, D.L. Massart, and Y. Vander Heyden, Development of a Ginkgo biloba fingerprint chromatogram with UV and evaporative light scattering detection and optimization of the evaporative light scattering detector operating conditions. *J. Chromatography A*, 2005b, **1085**(2), 230–239.

Vogt, F. and K. Booksh, Influence of wavelength-shifted calibration spectra on multivariate calibration models. *Appl. Spectroscopy*, 2004, **58**(5), 624–635.

Walczak, B. and D.L. Massart, Dealing with missing data, Part I. *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**(1), 15–27.

Wang, C.P. and T.L. Isenhour, Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal. Chem.*, 1987, **59**(4), 649–654.

Wang, Z.D., M. Fingas, S. Blenkinsopp, G. Sergy, M. Landriault, L. Sigouin, J. Foght, K. Semple, and D.W.S. Westlake, Comparison of oil com-

position changes due to biodegradation and physical weathering in different oils. *J. Chromatography A*, 1998, **809**(1–2), 89–107.

Wang, Z.D., M. Fingas, and K. Li, Fractionation of a light crude-oil and identification and quantitation of aliphatic, aromatic, and biomarker compounds by Gc-Fid and Gc-Ms, 1. *J. Chromatographic Sci.*, 1994a, **32**(9), 361–366.

Wang, Z.D., M. Fingas, and K. Li, Fractionation of a light crude-oil and identification and quantitation of aliphatic, aromatic, and biomarker compounds by Gc-Fid and Gc-Ms, 2. *J. Chromatographic Sci.*, 1994b, **32**(9), 367–382.

Wang, Z.D., M. Fingas, E.H. Owens, L. Sigouin, and C.E. Brown, Long-term fate and persistence of the spilled Metula oil in a marine salt marsh environment — degradation of petroleum biomarkers. *J. Chromatography A*, 2001, **926**(2), 275–290.

Wang, Z.D., M. Fingas, and D.S. Page, Oil spill identification. *J. Chromatography A*, 1999, **843**(1–2), 369–411.

Wang, Z.D., M. Fingas, and L. Sigouin, Using multiple criteria for fingerprinting unknown oil samples having very similar chemical composition. *Environ. Forensics*, 2002, **3**(3–4), 251–262.

Wang, Z.D., M. Fingas, and L. Sigouin, Characterization and source identification of an unknown spilled oil using fingerprinting techniques by GC-MS and GC-FID. *Lc Gc North America*, 2000, **18**(10), 1058.

Willse, A., A.M. Belcher, G. Preti, J.H. Wahl, M. Thresher, P. Yang, K. Yamazaki, and G.K. Beauchamp, Identification of major histocompatibility complex-regulated body odorants by statistical analysis of a comparative gas chromatography/mass spectrometry experiment. *Anal. Chem.*, 2005, **77**, 2348–2361.

Witjes, H., M. Pepers, W.J. Melssen, and L.M.C. Buydens, Modelling phase shifts, peak shifts and peak width variations in spectral data sets: Its value in multivariate data analysis. *Analytica Chimica Acta*, 2001, **432**(1), 113–124.

Wold, S., K. Esbensen, and P. Geladi, Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**(1–3), 37–52.

Wong, J.W.H., C. Durante, and H.M. Cartwright, Application of fast Fourier transform cross-correlation for the alignment of chromatographic and spectral datasets. *Anal. Chem.*, 2005, **77**, 5655–5661.

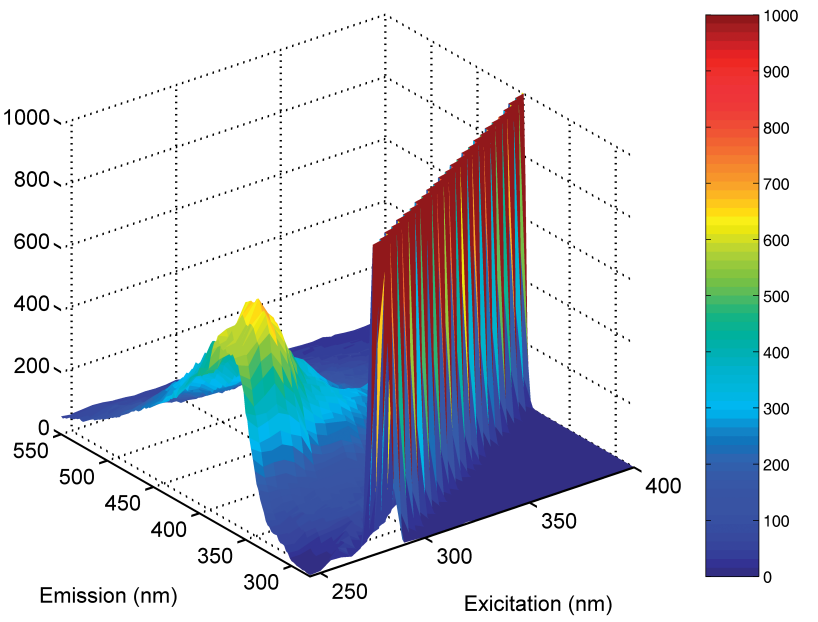

Figure 9-2 Fluorescence excitation-emission scans of an HFO. The vertical axis and scale bar show the fluorescence intensity. Modified from Christensen et al. (2005b).

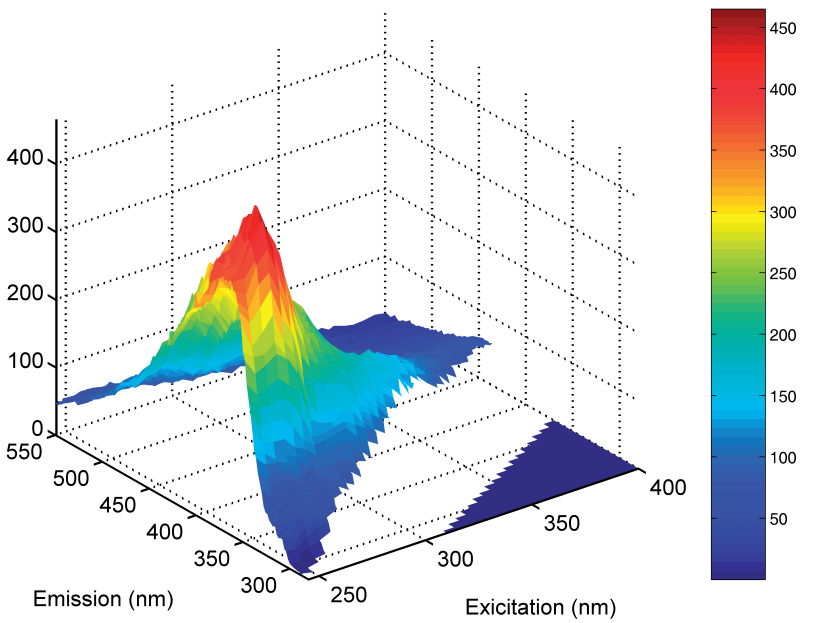

Figure 9-5 Preprocessed fluorescence EEM of the same oil sample as shown in Figure 9-2 after blank subtraction, insertion of missing values, and a small triangle of zeros and excitation/emission correction. The vertical axis and scale bar show the fluorescence intensity. Modified from Christensen et al. (2005b).