National Environmental Research Institute
Ministry of the Environment · Denmark

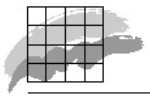
Chemometrics as a tool to analyse complex chemical mixtures

Environmental forensics and fate of oil spills

PhD Thesis

Jan H. Christensen

Roskilde University



National Environmental Research Institute
Ministry of the Environment · Denmark

Chemometrics as a tool to analyse complex chemical mixtures

Environmental forensics and fate of oil spills

PhD Thesis
2005

Jan H. Christensen

Roskilde University

Data sheet

Title:	Chemometrics as a tool to analyse complex chemical mixtures
Subtitle:	Environmental forensics and fate of oil spills. PhD thesis
Author:	Jan H. Christensen
Department:	Environmental Chemistry and Microbiology
University:	Roskilde University, Department of Life Sciences and Chemistry
Publisher:	National Environmental Research Institute © Ministry of the Environment
URL:	http://www.dmu.dk
Date of publication:	March 2005
Editing complete:	March 2005
Referee:	Niels Kroer, National Environmental Research Institute
Financial support:	Roskilde University, National Environmental Research Institute, Danish Natural Sciences Research Council
Please cite as:	Christensen, J.H., 2005. Chemometrics as a tool to analyse complex chemical mixtures – Environmental forensics and fate of oil spills. PhD thesis. Roskilde University, Department of Life Sciences and Chemistry and NERI, National Environmental Research Institute, Department of Environmental Chemistry and Microbiology. XX pages. http:/afhandlinger.dmu.dks

Abstract: Chemical characterisation of contaminant mixtures is important for environmental forensics and risk assessment. The great challenge in future research lies in developing suitable, rapid, reliable and objective methods for analysis of the composition of complex chemical mixtures. This thesis describes the development of such methods for assessing the identity (chemical fingerprinting) and fate (e.g. biodegradation) of petroleum hydrocarbon mixtures. The methods comply with the general concept that suitable methods must be rapid and inexpensive, objective with limited human intervention and at the same time must consider a substantial fraction of compounds in the complex mixture. A combination of a) limited sample preparation, b) rapid chemical screening analysis, c) fast and semi-automatic pre-processing, d) comprehensive multivariate statistical data analysis and e) objective data evaluation was used throughout the thesis.

Keywords:	Oil spills; petroleum hydrocarbons, complex mixtures, polyaromatic compounds, petroleum biomarkers, environmental forensics, chemical fingerprinting, biodegradation, weathering, correlation optimised warping, principal component analysis, parallel factor analysis, gas chromatography – mass spectrometry, fluorescence spectroscopy, data preprocessing
Cover photo:	Underwater photo of oil pollution at Bredemade Hage, Grønsund after the Baltic Carrier oil spill 29 March 2001
ISBN:	87-7772-860-2
Number of pages:	
Internet-version:	The report is available in electronic format from NERI's homepage http://www2.dmu.dk/1_viden/2_Publikationer/3_Ovrige/rapporter/phd_jch.pdf

Contents

Preface

This PhD thesis constitutes a fulfilment of the requirements for the Doctor of Philosophy degree at Roskilde University (RUC), Denmark. The thesis is comprised of an Introduction (including results and discussion sections, and conclusion and perspectives) and five papers, of which four have been published and one is submitted February 2005.

The work was done partly at the Department of Life sciences and Chemistry RUC, and partly at the Department of Environmental Chemistry and Microbiology, the National Environmental Research Institute (NERI). I would like to thank the two institutions for co-funding my work.

I greatly acknowledge my supervisors Asger B. Hansen (NERI) and John Mortensen and Ole Andersen (RUC) for inspiring supervision and their always positive and supportive attitudes. I remember especially the fishing trips to Roskilde Fjord (unfortunately too few) and the conferences we attended.

I am most indebted to my wonderful colleagues at the Department of Environmental Chemistry and Microbiology, NERI. It is because of them that my spirits were kept high through the difficult periods during the long hours in the laboratory and in front of the computer. I would also like to thank my project and master students through the years at Roskilde University. It has been inspiring working as supervisor for students that have interest in the same subjects as myself.

I am thankful to Lars Nørgaard and Rasmus Bro for allowing me to spend three inspiring months in the Chemometric group at the Royal Veterinary and Agricultural University during spring 2003. The large group of PhD students and scientists provided a stimulating environment for working with and discussing chemometrics. A special thank is expressed to Giorgio Tomasi for introducing me to "correlation optimised warping" and our fruitful corporation since then.

I also want to thank the people at Environment Canada in Ottawa, who I visited fourteen days in December 2004. Special thanks to my "Chinese parents" Zhendi and Amie for their great hospitality and kindness, and Bruce Holleborne for the many fruitful scientific discussions. I would also like to thank my PhD colleagues and friends Jesper Schmidt Hansen, Karin Liltorp, Karen Timmermann and Henrik Juhl Hansen for their interest in my work, their encouragement and support. Last but not least: I am most grateful to Lis for her understanding during periods with stress and long working hours away from home.

Roskilde, 1/3 2005

Jan H Christensen

List of papers

This thesis is based on the following papers:

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

(II) Christensen J.H., Hansen A.B., Tomasi G., Mortensen J. and Andersen O., Integrated Methodology for Forensic Oil Spill Identification, *Environmental Science & Technology*, **2004**, 38 (10), 2912-2918.

(III) Christensen J.H., Tomasi G. and Hansen A.B., Chemical Fingerprinting of Petroleum Biomarkers Using Time Warping and PCA, *Environmental Science & Technology*, **2005**, 39 (1), 255-260.

(IV) Christensen J.H., Hansen A.B., Mortensen J. and Andersen O., Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis, *Analytical Chemistry*, **2005**, 77 (7), 2210-2217.

(V) Christensen J.H., Hansen A.B., Karlson U., Mortensen J. and Andersen O., Multivariate Statistical Methods for Evaluating Biodegradation of Oil in the Environment, *Journal of Chromatography A*, **2005**, (Web Release Date: 1. August 2005).

Associated papers not part of the Ph.D. work is presented in appendix A-C.

(A) Vorkamp K., Christensen J.H., Glasius M. and Riget F.F., Persistent halogenated compounds in black guillemots (*Cepphus grylle*) from Greenland - levels, compound patterns and spatial trends, *Marine Pollution Bulletin*, **2004**, 48, 111-121.

(B) Vorkamp K., Christensen J.H. and Riget F.F., Polybrominated diphenyl ethers and organochlorine compounds in biota from the marine environment of East Greenland, *Science of the Total Environment*, **2004**, 331, 143-155.

(C) Glasius M., Christensen J.H., Platz J. and Vorkamp K., Halogenated organic contaminants in marine fish and mussels from southern Greenland – pilot study on relations to trophic levels and local sources, *Journal of Environmental Monitoring*, **2005**, 7, 127-131.

List of Abbreviations

BTEX	The collective name of benzene, toluene, ethylbenzene and *o*-, *m*- and *p*-xylene isomers
COW	Correlation optimised warping
DAS	Diasteranes
EEM	Excitation-emission matrices
GC-FID	Gas chromatography – flame ionisation detection
GC-GC	Two-dimensional gas chromatography
GC-IRMS	Gas chromatography – isotope ratio mass spectrometry
GC-MS	Gas chromatography – mass spectrometry
HFO	Heavy fuel oil
LC-DAD	Liquid chromatography – diode array detection
LC-MS	Liquid chromatography – mass spectrometry
LFO	Light fuel oil or diesel
Lub	Lubricating oil
MD	Methyldibenzothiophene
MF	Methylfluorene
MP	Methylphenanthrene
PAC	Polycyclic aromatic compound
PARAFAC	Parallel factor analysis
PC	Principal component
PCA	Principal component analysis
QA/QC	Quality assurance and quality control
RS	Rearranged steranes
RSD_A	Relative analytical standard deviation
RSD_S	Relative sampling standard deviation
SIM	Selected ion monitoring
TPH	Total petroleum hydrocarbon
WLS	Weighted least squares
WLS-PCA	Weighted least squares – PCA

Summary

Five novel methods for data pre-processing, chemical fingerprinting and evaluation of biodegradation of complex chemical mixtures were developed in this PhD-thesis and described in five international peer-reviewed papers. In Paper I *"Chromatographic preprocessing of GC-MS data for analysis of complex chemical mixtures"* published in Journal of Chromatography A (2005), a semi-automatic method for processing complex first-order chromatographic data, such as gas chromatography – mass spectrometry (GC-MS) with selected ion monitoring (SIM) was developed. The method is based on automated peak matching, initial parameterisation, alternating background noise reduction and peak estimation using mathematical functions with few parameters (i.e., three to four). It allows resolvation of convoluted peaks, and an exponential-Gaussian hybrid improves the description of asymmetric peaks. Chromatograms are converted into semi-quantitative variables (e.g., diagnostic ratios) which enable a comprehensive analysis of complex chemical mixtures. The method was tested on chromatographic data from twenty replicate oil samples, and it was observed to be less time consuming and more objective compared to commercial software, while retaining comparable data quality.

Paper II *"Integrated Methodology for Forensic Oil Spill Identification"* published in Environmental Science & Technology (2004) presents an integrated method for forensic oil spill identification. The method is comprised of chromatographic data processing, variable-outlier detection, principal component analysis (PCA), estimation of analytical and sampling uncertainties and statistical evaluation. The pre-processing method described in Paper I was used for peak quantification and calculation of diagnostic ratios and their uncertainties. The use of PCA enabled simultaneous analysis of a large number of diagnostic ratios. Weathering was taken into account by considering the sampling uncertainties estimated from replicate spill samples. Statistical evaluation based on the null-hypothesis of scores ensured an objective matching of oil spill samples with suspected source oils, and classification into positive match, probable match and non-match. The data analysis can be refined if two or more source oils are classified as probable match by using weighted least squares fitting of the principal components (WLS-PCA) and by focussing the PCA on a subset of source oils including all the probable matches (local PCA models). The method was tested on four groups of diagnostic ratios derived from petroleum biomarkers (terpanes and steranes) and ratios within homologue series of polyaromatic compounds (PACs). The sources of two spill samples (Norwegian crude oils from Oseberg East and Oseberg Field Centre) were identified and distinguished from the closely related source oil (crude oil from Oseberg South East).

Paper III *"Chemical Fingerprinting of Petroleum Biomarkers Using Time Warping and PCA"* published in Environmental Science & Technology (2005) presents a novel method on chemical fingerprinting of petro-

leum biomarkers. The method allowed for analyses of chromatograms using a fast and highly objective procedure comprised of GC-MS analysis, pre-processing of GC-MS (SIM) chromatograms and PCA of selected regions. Once the PCA model is constructed, the complete data analysis of a new oil sample requires only a short period of time (a few seconds). Furthermore, as long as the variation between oils in the calibration set is sufficient, the PCA can distinguish coeluting peaks, which is far more difficult when relying on peak quantification as in Paper II. The pre-processing consists of baseline removal by calculating the derivative, normalisation and chromatographic peak alignment using correlation optimised warping (COW). The method was applied to chromatograms of m/z 217 (tricyclic and tetracyclic steranes). 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 was analysed. Four reliable components could be extracted from data, whereas the chemical information in additional components was confounded with the residual misalignment. The four principal components were interpreted as: boiling point range (PC1), clay content (PC2) and carbon number distribution of sterols in the source rock (PC3), and thermal maturity of the oil (PC4). The *Baltic Carrier* spill samples were clustered in principal components 1 to 4 together with oil samples from the tank of the *Baltic Carrier* (source oil). Deselecting the most uncertain variables or using WLS-PCA enhanced the resolution power.

Paper IV *"Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis"* published in Analytical Chemistry (2005), presents a novel approach for matching oil samples by fluorescence spectroscopy combined with three-way decomposition of spectra. It offers an objective fingerprinting based on the relative composition of PACs in oils and is complementary to gas chromatography – flame ionisation detection (GC-FID) for initial screening of oil samples. Portable fluorescence spectrometers are commercially available, and combined with a laptop, the procedure may be implemented for on-site (e.g., onboard ships during state-port control) initial prescreening. Parallel factor analysis (PARAFAC) was applied to fluorescence excitation-emission matrices (EEM) of heavy fuel oils (HFO), light fuel oils (LFOs), lubricating oils (Lubs), crude oils, unknown oils and a sample collected in the spill area two weeks after the *Baltic Carrier* oil spill (Denmark, 2001). EEMs (112 in total) were decomposed into a five-factor PARAFAC model using excitation wavelengths from 245 to 400 nm and emission wavelengths from 280 to 550 nm. Comparison of PARAFAC factors with EEMs of PAC standards with two to five rings indicated that each of the factors could be related to a mixture of oil-characteristic PACs with similar fluorescence: A mixture of naphthalenes and dibenzothiophenes, fluorenes, phenanthrenes, chrysenes and five-ring PACs. Oils were grouped in score plots according to oil type, and except for HFOs and crude oils, the method easily discriminated between the four types. The spill sample was correctly assigned as a HFO with similar PAC pattern as oil from the cargo tank of the *Baltic Carrier* by comparing the correlation coefficient of scores for the oil spill sample and possible source oils (i.e., oils in the database).

Paper V *"Multivariate Statistical Methods for Evaluating Biodegradation of Oil in the Environment"* submitted to Journal of Chromatography A (February 2005) presents methods for evaluating natural attenuation and bioremediation of oil in the environment. The methods are based on PCA or WLS-PCA of aligned and normalised regions of GC-MS chromatograms and diagnostic ratios of individual isomers of PACs. As opposed to the method in Paper III, the normalisation of chromatograms was performed after retention time alignment. The optimal combination of pre-processing and data analysis was determined by optimising the alignment of replicate reference samples and by minimising the variability of duplicate biodegradation samples in the PCA. The methods were applied to data from an *in vitro* biodegradation experiment. A North Sea crude oil was exposed to three mixtures of bacterial strains over a 1-year period with five sampling times. The variation in degradability within groups of isomers of methylfluorene (m/z 180), methylphenanthrene (m/z 192) and methyldibenzothiophene (m/z 198) were used to evaluate the effects of microbial degradation on the oil composition. Principal component 1 (PC1) described the general changes in the isomer distribution due to microbial degradation. The sterile control samples and samples unaffected by biodegradation were clustered together with oil samples from day 0, at low PC1, and degraded samples were located at increasing PC1 scores. In addition M samples were separated from U and R samples along PC2. These observations demonstrated that the strain mixture (M), which consisted of R (alkane degraders and surfactant producers) and U strains (PAC degraders), affected the PAC isomer patterns differently than the U- and R-strains independently.

The general methodologies for assessment of petroleum hydrocarbons in the environment presented in the five papers consist of:

- Fast sample preparation and comprehensive chemical analysis based on semi-quantitative methods
- Semi-automatic data pre-processing for reducing information unrelated to the chemical variation. Variable selection or scaling were performed to focus the subsequent data analysis on the variations with highest diagnostic power
- Two- and multiway statistical decomposition of data using respectively PCA and WLS-PCA and PARAFAC
- Data evaluation by visual inspection of score and loading plots and more objective comparisons using similarity indices and statistical tests

The methods do not only possess a great potential for oil spill identification and for the study of oil biodegradation, but also for more general purposes. Hence, the developed methods can be used for chemical fingerprinting, assessment of fate and monitoring of complex chemical mixtures in general such as persistent organic pollutants (e.g., polychlorinated biphenyls and brominated flame retardants. The combination of compound concentrations, diagnostic ratios and PCA have been used in the associated papers A, B and C to analyse the pollution pattern in biota (e.g., marine fish, mussels, black guillemots and other biota) from the Greenland environment.

Sammenfatning

Metoder til brug ved dataforbehandling, kemisk fingerprinting og analyse af bionedbrydning af komplekse kemiske blandinger er beskrevet i fem videnskabelige artikler. I artikel I *"Chromatographic preprocessing of GC-MS data for analysis of complex chemical mixtures"* publiceret i Journal of Chromatography A (2005) er der udviklet en delvist automatiseret metode til forbehandling af komplekse førsteordens kromatografiske data. Metoden kan anvendes til forbehandling af data, inden den kemometriske analyse påbegyndes, og den er baseret på automatisk identifikation af kromatografiske toppe, initial estimering af top parametre samt alternerende reduktion af baggrund og top estimering ved brug af matematiske funktioner med få parametre (3-4). Metoden giver mulighed for at modellere overlappende kromatografiske toppe samt forbedre beskrivelsen af asymmetriske toppe. I metoden omdannes kromatogrammer til semi-kvantitative variable (diagnostiske forhold), hvilket gør det praktisk muligt at udføre omfattende analyse af komplekse kemiske blandinger. Metoden er testet på kromatografiske data fra tyve replikate olieprøver, og resultaterne er derefter sammenlignet med resultater fremkommet ved brug af kommercielt GC-MS software. Resultaterne viser, at den nyudviklede metode er langt mindre tidskrævende og mere objektiv end standard metoder, men med sammenlignelig data kvalitet.

Artikel II *"Integrated Methodology for Forensic Oil Spill Identification"* publiceret i Environmental Science & Technology, omhandler en integreret metode til identifikation af oliespild. Metoden består af kromatografisk dataforbehandling, variable-outlier detektion, principal komponent analyse (PCA), estimering af analyse- eller prøvetagningsusikkerheder samt statistisk evaluering. Forbehandlingsmetoden, beskrevet i artikel I, er anvendt til identifikation og kvantificering af kromatografiske toppe samt beregning af diagnostiske forhold og usikkerheder. Anvendelse af PCA gjorde det muligt at analysere et stort antal diagnostiske forhold simultant. Oliens forvitring i miljøet blev taget i betragtning ved at anvende prøvetagningsusikkerhederne, estimeret fra replikate spildprøver, til de statistiske tests. Statistisk evaluering baseret på nulhypotesen for "score" værdier sikrede en objektiv sammenligning af spildprøver med olie fra mistænkte kilder, og klassificering i "positiv match", "mulig match" eller "ingen match". Når mere end én olieprøve blev klassificeret som "mulig match", blev dataanalysen forfinet ved anvendelse af vægtede-mindste-kvadraters estimering af de principale komponenter (WLS-PCA) eller ved at fokusere dataanalysen på et delsæt af olier, inklusiv alle de mulige kilder. Metoden blev testet på fire grupper af diagnostiske forhold beregnet ud fra biomarkører (terpaner og steraner) og forhold beregnet mellem stoffer indenfor homologe serier af PACer. Kilden til to spildprøver (norske råolier fra Oseberg Øst og Oseberg Field Centre) blev identificeret og adskilt fra råolie fra Oseberg Sydøst.

Artikel III *"Chemical Fingerprinting of Petroleum Biomarkers Using Time Warping and PCA"* publiceret i Environmental Science & Technology

(2005) præsenterer en metode til kemisk fingerprinting af biomarkører i olie. Metoden er hurtig og objektiv og består af GC-MS analyse, dataforbehandling af GC-MS (SIM) kromatogrammer og PCA af udvalgte udsnit. Dataanalyse og identifikation af en ny ukendt olieprøve kræver kun få sekunder, når først PCA modellen er opbygget. Sålænge variationen mellem olier i kalibreringssættet er tilstrækkelig, kan PCA adskille overlappende toppe, hvilket er langt mere vanskeligt, når forbehandlingen indbefatter identifikation og kvantificering af kromatografiske toppe, som er tilfældet i artikel II. Forbehandlingen består af fjernelse af basislinjen ved at udregne den første afledte, normalisering og kromatografisk synkronisering ved brug af "correlation optimised warping" (COW). Metoden er anvendt til databehandling af m/z 217 ionsporet (tricykliske og tetracykliske steraner). Fire PCA komponenter blev fundet, mens den kemiske information i højere komponenter var blandet med variation forårsaget af utilstrækkelig synkronisering. De principale komponenter blev fortolket som: kogepunktsområde for raffinerede råolieprodukter (PC1), lerindhold i kildebjergarten (PC2), fordelingen af kulstof (antal) i steroler i kildebjergarten (PC3) og oliens termiske modenhed (PC4). Spildsprøverne, indsamlet fra kystmiljøet ved Grønsund i ugerne efter *Baltic Carrier* oliespildet, var grupperet i PC1 til PC4 sammen med olieprøver fra lasttanken på *Baltic Carrier*. Variable selektion eller WLS-PCA forbedrede modellens evne til at adskille prøver fra forskellige kilder (forbedret resolutions evne).

Artikel IV *"Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis"* publiceret i Analytical Chemistry (2005) præsenterer en ny fremgangsmåde til identifikation af olieprøver ved fluorescence spektroskopi kombineret med trevejs-dekomponering af spektre. Metoden tilbyder en objektiv fingerprinting baseret på den relative sammensætning af PACer i olie, og er komplementær til GC-FID til screening af olieprøver. Transportable fluorescence spektrometerer er kommercielt tilgængelige, og kombineret med en laptop computer kan metoden implementeres onsite (f.eks. ombord på skibe ved "state-port kontrol") til pre-screening af spildprøver. Fluorescence excitation-emission matricer (EEMer) af tunge fuel olier (HFO), dieselolier, smøreolier, råolier, ukendte olier og en prøve indsamlet i spildområdet to uger efter *Baltic Carrier* oliespildet (Danmark 2001) blev analyseret ved brug af Parallel factor analyse (PARAFAC). EEMer (i alt 112) blev dekomponeret til en femfaktor PARAFAC model ved brug af eksitations bølgelængder fra 245 til 400 og emissions bølgelængder fra 280 til 550 nm. Desuden blev EEMer af PAC standarder med to til fem benzenringe sammenlignet med PARAFAC faktorerne. Resultaterne indikerer, at hver af faktorerne kan relateres til blandinger af oliekarakteristiske PACer med ensartet fluorescence: henholdsvis en blanding af naftalener og dibenzothiofener, fluorener, phenanthrener, chrysener og fem-rings PACer. Olierne grupperer sig i score plottene afhængig af olietype, og undtagen for HFOer og råolier kan metoden diskriminere mellem de fire typer. Spildprøven er, ved at sammenligne korrelationskoefficienten af scores fra oliespildsprøven og mulige kilder, korrekt identificeret som en HFO med ensartet PAC mønster som olie fra lasttanken på *Baltic Carrier*.

Artikel V *"Multivariate Statistical Methods for Evaluating Biodegradation of Oil in the Environment"* submitted til Journal of Chromatography A (Februar 2005) præsenterer to metoder til vurdering af den naturlige nedbrydning og bioremediering af olie i miljøet. Metoderne er baseret på PCA eller WLS-PCA af synkroniserede og normaliserede udsnit af GC-MS kromatogrammer, samt diagnostiske forhold af individuelle PAC isomerer. Normalisering af kromatogrammerne er, i modsætning til artikel III, udført efter synkroniseringen. Den optimale kombination af forbehandling og dataanalyse er fundet ved henholdsvis at bestemme den optimale synkronisering af replikate referenceprøver samt ved minimering af variabiliteten af duplikate bionedbrydningsprøver i PCA. Metoderne er anvendt på data fra et *in vitro* bionedbrydningsforsøg. En Nordsøolie blev eksponeret til tre forskellige blandinger af bakterier over en tidsperiode på 1 år med fem indsamlingstidspunkter. Selektiv nedbrydning af isomerer indenfor methylfluorener (m/z 180), methylphenanthrener (m/z 192) og methyldibenzothiophener (m/z 198) er brugt til at evaluere effekten af mikrobiel nedbrydning på oliens sammensætning. Principal komponent 1 (PC1) beskriver de generelle ændringer i isomersammensætningen forårsaget af mikrobiel nedbrydning. Steril kontrollerne og de upåvirkede prøver er grupperet ved lav PC1 sammen med olieprøver fra dag 0, mens nedbrudte prøver har højere PC1 score værdier. Endvidere adskiller PC2 M-prøverne fra U- og R-prøverne. Dette demonstrerer at stammeblandingen (M-prøver), som består af R (alkannedbryderer der danner overflade aktive stoffer) og U-prøverne (PAC nedbryderer), påvirkede PAC isomermønstrene anderledes end, hvad U- og R-stammerne gjorde individuelt.

Den generelle metode til vurdering af olie i miljøet præsenteret i de fem artikler består af :

- Begrænset prøveforberedelse og hurtig og omfattende kemisk analyse baseret på semi-kvantitative metoder
- Dataforbehandling ved at reducere den information, som ikke er relateret til den kemiske variation. Variabel selektion og skalering er udført for at fokusere dataanalysen på data med højest diagnostisk styrke
- To- og multivejs statistisk dekomponering af data ved brug af henholdsvis PCA og WLS-PCA samt PARARAC analyse.
- Data evaluering ved visuel inspektion af score og loading plots samt mere objektive sammenligninger ved brug af similaritets indekser og statistiske tests

Metoderne har ikke kun et stort potentiale i forbindelse med identifikation af oliespild og analyse af bionedbrydning af olie, men også til mere generelle formål. De udviklede metoder kan således anvendes ved kemisk fingerprinting, undersøgelse af skæbne og monitering af andre komplekse kemiske blandinger, såsom persistente organiske miljøfremmede stoffer (polychlorede biphenyler og bromerede flammehæmmere). Kombinationen af stofkoncentrationer, diagnostiske forhold og PCA har i de associerede artikler A, B og C være anvendt til analyse af forureningsmønstre i biota (marine fisk, muslinger, tejst og andre biota) fra det Grønlandske miljø.

1 Introduction

Thousands of chemicals with different physicochemical properties and toxicity are used in our society and many of them can be detected in environmental (e.g., water, sediment, soil, air and biota) and human samples. Yet risk assessment is often limited to one chemical at a time and environmental monitoring to very specific and tiered analyses of relatively few compounds. Standard methods applied for environmental chemical analysis such as gas chromatography – mass spectrometry (GC-MS) and liquid chromatography – mass spectrometry (LC-MS) can generate extensive amounts of compound-specific data. The limiting factor for exploiting the full potential of hyphenated analytical techniques lies in the lack of appropriate tools to process, analyse and evaluate large data sets comprising measurements of hundreds or thousands of single compounds. Furthermore, the resources required to assess the full sample complexity using specific and tiered chemical analysis for each subgroup of analytes are enormous.

This PhD-thesis describes the development and application of general methods for the analysis of complex chemical mixtures. It complies with the present limitations in methods for chemical analysis and assessment of complex mixtures. The overall methodology comprises fast and easy-to-run chemical analysis (semi-quantitative analyses), fast data pre-processing with limited human intervention, fast and comprehensive data analysis using multivariate statistical methods and objective data evaluation. Crude oil and petroleum products are used as, general case examples of complex chemical mixtures throughout the thesis. However, the chemometric tools may also be used to analyse other types of complex chemical mixtures. In the following, a brief overview of petroleum as a class of inherently complex and ubiquitously distributed mixtures will be given.

Crude oil and petroleum products

Crude oil and refined petroleum products play an important role in the modern society. Petroleum products are used as fuels in cars (gasoline and diesel), aircrafts (jet fuels) and ships (heavy fuel oil); for heating and electricity generation; as lubricants in machinery; as asphalt for road building and for the production of chemicals and plastics. The estimated oil production was 75.34 million barrels/day in 2001 (CIA World Fact Book 2004).

The exploration, production, transportation and widespread use of crude oil and petroleum products inevitably result in intentional and accidental releases to the environment. Waterborne oil spills of unknown origin often occur in rivers, in open water and in coastal waterways, and they range from the continuous leakage from land sources and illegal tank washings at sea to larger spill accidents such as the 2002 *Prestige* oil spill. World-wide, the number of spills is enormous. In Danish maritime territory alone, the frequency of minor spills (e.g., due to tank washings) was approximately 400 per year during a 15 year period from 1987 to 2001.

Oil generation

Petroleum is a complex mixture of hydrocarbons that exist naturally in gaseous (natural gas), liquid (crude oil), and solid (asphalt) states. It is derived from a variety of biochemical compounds present in algae, bacteria, phyto- and zooplankton and higher plants. The conversion and transformation of biochemical compounds (e.g., lipids, fatty acids, proteins, porphyrins, sterols and terpenoids) into petroleum is completed through a selection of maturation processes: diagenesis, catagenesis and metagenesis in time scales of hundreds of millions of years (1). The composition of crude oils can vary substantially depending on the starting organic materials and the maturation history in the source rock and the reservoir.

Oil composition

Petroleum hydrocarbons can be broadly divided into saturated, aromatic and polar compounds. Saturates are the predominant class of hydrocarbons in most crude oils and include straight and branched chain saturates (i.e., paraffins) and cycloalkanes (naphthenes). The aromatics are cyclic, planar compounds that resemble benzene in electronic configuration and chemical behaviour. The aromatics are mainly comprised of BTEX (the collective name of benzene, toluene, ethylbenzene, and *o*-, *m*- and *p*-xylene isomers), other alkyl-substituted benzene compounds such as naphthenoaromatics and polycyclic aromatic compounds (PACs) including the alkylated C_0 - C_4-PACs that are characteristic in oil. The polar fraction includes heterocyclic hydrocarbons such as nitrogen, oxygen and sulphur containing PACs, phenols, acids and alcohols.

Oil toxicity

Oil spills can affect ecosystems and human health due to the content of toxic and mutagenic compounds. BTEXs have acute and long-term toxic effects, and benzene is a recognised carcinogen**.** PACs, present in relative high concentrations in most oils, form a large group of relatively persistent compounds, several being carcinogenic and/or mutagenic. Accordingly, the US Environmental Protection Agency (http://www.epa.gov) has classified 16 individual PACs as priority pollutants.

Analytical techniques

A wide selection of analytical techniques has been used for analysis of petroleum hydrocarbons in environmental samples following oil spills (petrogenic hydrocarbons) and emission/deposition following fossil fuel combustion (pyrogenic hydrocarbons). The conventional methods include gravimetric determination, infrared spectroscopy, ultraviolet spectroscopy, fluorescence spectroscopy (2,3), thin layer chromatography, high performance liquid chromatography, gas chromatography with flame ionisation detection (GC-FID) (4), GC-MS (5,6), two-dimensional gas chromatography (GC-GC) (7,8), gas chromatography-isotope ratio mass spectrometry (GC-IRMS) (9,10) and GC-MS-MS (11).

The standard method for chemical characterisation of oil consists of initial screening using GC-FID (4) or fluorescence spectroscopy (2,3) followed by a more comprehensive analysis using GC-MS with electron impact ionisation in selected ion monitoring (SIM) mode (5,6). GC-MS can resolve a broad range of petroleum hydrocarbons important for comprehensive characterisation of oil, which include petroleum biomarkers (e.g., steranes, terpanes and sesquiterpanes) and

PACs. GC-MS-MS and GC-GC are other instrumental techniques capable of providing compound-specific information for oil analysis (8,11).

Hydrocarbon fingerprinting

Chemical fingerprinting of petroleum hydrocarbons was developed by the petroleum industry to understand and track the source of crude oils and natural gases. In environmental chemistry and environmental forensics, methods basically similar to those of petroleum geochemistry are applied to reveal the history and source of environmental pollutants. Environmental forensics is defined as a scientific methodology for identifying environmental contaminants and for determining their sources and time of release. Chemical fingerprinting can be defined as a set of techniques applicable in the environmental assessment of complex mixtures, like oil and petroleum products. Chemical fingerprinting includes characterising (i.e., fuel type), quantifying (i.e., concentration), differentiating (i.e., from a similar source of pollution) and identifying (i.e., trace identity and fate) a spill sample based on its chemical composition (i.e., distribution patterns of different compounds).

More specifically, hydrocarbon fingerprinting is important to:

- Defensibly determine oil source(s) and distinguish spilled oil from background hydrocarbons of biogenic and pyrogenic origin
- Determine the fate of oil. Crude oil and refined petroleum products released into the environment are subject to numerous transport and transformation processes commonly known as weathering processes.
- Monitor petroleum hydrocarbons in compartments of the environment with the aim of assessing the impact on the ecosystem (risk assessment)

Data analysis

Data analysis is an important part of the assessment of oil in the environment after oil spills whether on land or water. The standard methods are based on comparison of bulk oil properties such as total petroleum hydrocarbon concentration (TPH), visual comparison of fluorescence spectra, GC-FID or GC-MS chromatograms, concentrations of source-specific markers, bar plots of the distribution of oil-characteristic PACs and lists and double plots of diagnostic ratios. Thus, and despite the inherent complexity of crude oil and the fact that GC-MS can generate extensive amounts of compound-specific data, chemical characterisations are mostly limited to few selected compounds or a visual comparison of data.

Weathering processes

Weathering processes can be divided into physical (e.g., evaporation, emulsification, natural dispersion, dissolution and sorption), chemical (photodegradation) and biological processes (microbial degradation) (12). Oil components are redistributed in compartments of the environment (e.g., air, water, sediment, soil and biota) during physical weathering and transformed during chemical and biological weathering processes. Evaporation and dissolution and, depending on the meteorological conditions, UV degradation are the most im-

portant weathering processes immediately (hours - few days) after oil spills on water. In contrast, biodegradation affects the long-term (months – years) fate of oil.

Numerous authors have investigated microbial degradation of mixtures of petroleum hydrocarbons; *in situ* (13-16) and under laboratory conditions (17-20). The general biodegradation trend observed in these investigations is that enhanced molecular complexity leads to a decrease in the susceptibility towards microbial attack. Specifically, the rate of PAC degradation in the environment decreases with increasing number of rings and the degree of alkylation (20-22). The general degradation order of alkylated PACs is $C_0 > C_1 > C_2 > C_3 > C_4$, where C_x denotes the total number x of carbon atoms of the substituents. These effects correspond to those of physical weathering processes such as evaporation and dissolution, since the physicochemical properties of PACs (i.e., boiling point and solubility) also depend on ring size and degree of alkylation (20).

Evaluation methods

Bulk properties such as TPH, measured by GC-FID, and gravimetric analysis of the aliphatic, aromatic and polar oil fractions have been used frequently for chemical fingerprinting and for evaluating the effects of oil weathering (13,23,24). Furthermore, GC-FID is often used for screening analysis of mainly *n*-alkanes and PACs with 2-3 rings. Subjective pattern matching of GC-FID fingerprints were used to assess the changes in composition of the *Amoco Cadiz* oil 13 years after the spill (13). GC-FID fingerprints were also used to classify sediment samples, collected 25 years after the *Nipisi* land spill, according to the levels of oil contamination and the extent of weathering. Sediment samples were classified into: a) background samples, b) highly weathered samples, c) lightly to moderately weathered and d) lightly contaminated with oil and vegetation hydrocarbons (16).

The heterogeneity encountered in field samples can be corrected by normalisation to a conservative internal marker compound such as 17α, 21β-hopane (25-28), 17α, 21β-norhopane (24) or vanadium (29). These methods have been used frequently for chemical fingerprinting in weathering studies and to calculate the percent loss of oil or individual analytes (22). Univariate plots of percent loss based on preserved internal markers describe the combined effects of physical, chemical and biological weathering processes. The isolated effects of microbial degradation can only be described in controlled laboratory experiments by subtracting the loss in sterile controls. Conversely, low-molecular-weight hydrocarbons, such as heptadecane, octadecane, pristane, phytane and 2-3 ring PACs, may in field samples be heavily affected by physical weathering processes.

Several authors have observed that microbial degradation is isomer specific (16,18,20,30,31). Changes in heptadecane/pristane and octadecane/phytane have long been used as indicators of biodegradation (15,32). Likewise, preferential degradation of specific isomers within homologue PAC series have been described since the 1980s (18,31,33). Volkman (1984) found that dimethylnaphthalenes (DMN) were degraded at different rates, with isomers having β-methyl substituents being most susceptible to microbial attack (33). Rowland *et al.* (1986)

observed preferential degradation of 2,7-DMN relative to 2,6-DMN, and that 9-methylphenanthrene was more persistent than the other isomers, in both field and laboratory samples (18). Recently, Wang *et al.* found preferential degradation within alkylated PAC families of C_1-C_3-naphthalenes, methylfluorenes (MF), methylphenanthrenes (MP) and methyldibenzothiophenes (MD) by subjective pattern matching and univariate comparison of diagnostic ratios of individual isomers (16,20,34). The studies by Wang *et al.* included field samples from the *Nipisi* oil spill (16) and laboratory biodegradation tests using a standard freshwater inoculum comprised of six microbial strains (three aliphatic and three aromatic degraders) (20,34). In particular, Wang *et al.* (1998) observed isomer specific microbial degradation as a decrease in the ratios (3MP+2MP) / (4+9MP+1MP), (2+3MD/1MD) and ∑MF/1MF) (16,20), where xMP for example denotes the position x of the methyl substituent in the phenanthrene skeleton.

Changes in isomer patterns within alkylated PAC series, and the corresponding diagnostic ratios, are highly specific for biodegradation as no such changes occur during physical weathering (20). In contrast, a few studies have shown that chemical weathering can lead to changes in the isomer patterns, although the sequences of alteration are different compared to those observed during biodegradation (35). Jacquot *et al.* (1996) observed the sequence 2MP < 1MP < 3MP < 4+9MP of increasing photodegradability for methylphenanthrenes. Photodegradation led to an increase in the (3MP+2MP) / (4+9MP+1MP) ratio. The differences in preferential degradation within homologue PAC series during biodegradation and photodegradation may be used to distinguish between chemical and biological weathering effects.

Isomer distributions within PAC homologue series have also been used frequently for chemical fingerprinting since the distribution is unaffected by short-term weathering processes. Wang *et al.* (1995) used methyldibenzothiophenes as markers for differentiation and source identification of crude and weathered oils (34). Petroleum biomarkers, such as terpanes and steranes, are however, the most important and frequently used group for petroleum hydrocarbon fingerprinting. Biomarkers are complex molecules derived from the ancient biota forming petroleum products. They show little or no structural changes compared to the parent organic molecules, so-called biogenic precursors, e.g., hopanoids, sterols and steroids. Thus, biomarkers are especially useful compounds for chemical fingerprinting since the types and relative amounts vary greatly depending on the starting organic materials and the maturation history in the source rock and the oil reservoir. Biomarker diagnostic parameters have long been recognised and are widely used by geochemists for oil correlation (oil-source rock and oil-oil correlation); determination of organic input and precursors; depositional environment; assessment of thermal maturity; evaluation of oil in-reservoir biodegradation and for oil spill identification purposes (6,36-40).

Multivariate statistical methods

Multivariate statistical methods have been used for data analysis in organic geochemistry since the 1980s (41,42). The methods have only

recently been adopted for oil spill identification (43-48) and for studying the fate of petroleum hydrocarbons in the environment (49,50).

Lavine *et al.* (2001) employed pattern recognition and principal components analysis (PCA) to study spilled jet fuels which had undergone weathering in a subsurface environment, and it enabled them to classify these fuels into five types (46). Aboul-Kassim and Simoneit used a variety of statistical techniques for source oil identification (43,44). In their analysis of the aliphatic and aromatic composition in particulate fallout samples in Alexandria (43), multivariate statistical analyses, including extended Q-mode factor analysis and linear programming, were performed in order to reduce the hydrocarbon data set into a meaningful number of end members (sources). Their analysis indicated that there were two significant end members explaining 90% of the total variation among samples and confirming petrochemical (79.6%) and thermogenic/pyrolytic (10.4%) sources in the model. In a study of sediment samples in the Eastern Harbour of Alexandria (44), a similar multivariate statistical approach, including factor analysis and linear programming techniques, was used to determine the end member compositions and evaluate sediment partitioning and transport in the Eastern Harbour area. The authors found that untreated sewage was the main source of petroleum hydrocarbons in the Eastern Harbour area rather than direct inputs from boating activities or urban run-off. Stout *et al.* (2001) analysed a suite of diagnostic PAC and biomarker ratios with PCA (48). 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.

Burns *et al.* (1997) used PCA and a least-squares iterative matching procedure to allocate PACs in intertidal and subtidal sediment samples from the Prince William Sound of Alaska to 30 potential sources (45). They used PCA 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. Subsequently, the least-squares model was used to estimate the source mix, with the best least-squares fit of 36 PAC analytes including the parent and alkylated homologues of naphthalene, phenanthrene, fluorene, dibenzothiophene and chrysene. Isomers were grouped by number of carbons in alkyl substituents and PAC family and treated as individual analytes. In a 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 (47). 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. The analysis suggested mixed sources whose contributions varied significantly across the sampling area.

1.1 Aim of study

The overall aim of this study has been to develop new and improved chemometric based tools for analysis of complex chemical mixtures. More specifically, the tools were developed for environmental forensics and for studying the fate of oil spills. The concept has been to reduce the time and costs required for analysis, while at the same time improve the objectivity and increase the number of individual compounds considered. The research strategy is illustrated in Figure 1. It comprises four steps: chemical analysis, data pre-processing, data analysis using chemometrics and data evaluation. The included papers (Paper I – V) focus on novel methods for data pre-processing, environmental forensics and studying the fate of oil spills.

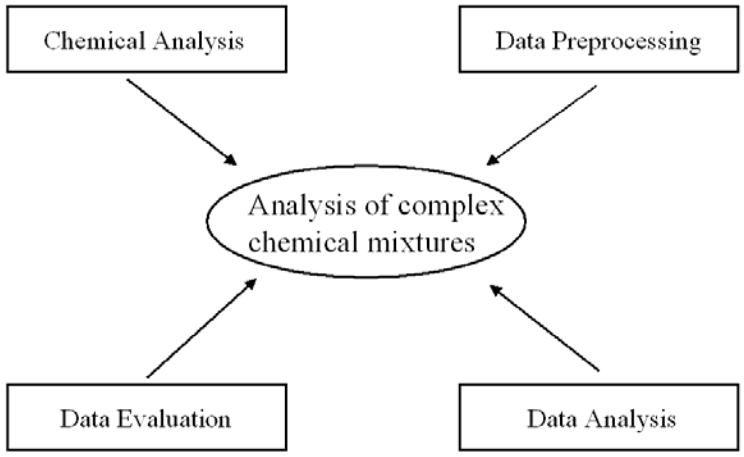


Figure 1. The research strategy used in this study is based on four steps: chemical analysis, data pre-processing, analysis and evaluation. The aim has been to develop rapid, reliable and objective tools for comprehensive analysis and characterisation of complex chemical mixtures.

2 Results and discussion

The following sections present and discuss the overall philosophy and results obtained in the submitted papers. The general protocol used for chemical fingerprinting in Paper II-IV and for the study of fate of oil in the environment (Paper V) is illustrated in Figure 2.

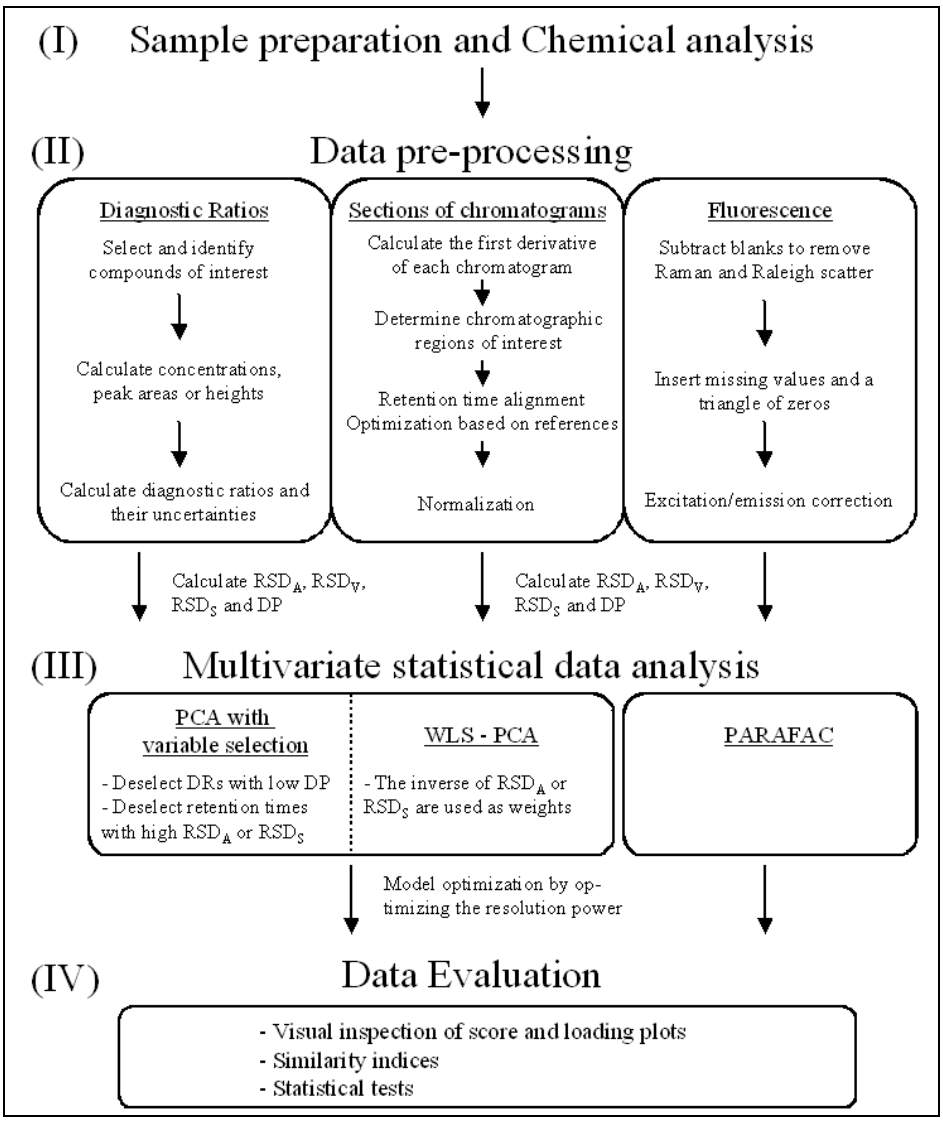


Figure 2. Flowchart for the general methodology for analysis of complex chemical mixtures with respect to environmental forensics and the study of the fate of spilled oil in the environment.

Procedures for data import, pre-processing, analysis and evaluation have been implemented in Delphi 4.0 (Borland) and Matlab 6.5 (The Mathworks). The procedure for chromatographic pre-processing of GC-MS data for analysis of complex chemical mixtures (Paper I) was implemented in Borland Delphi 4.0 object oriented programming, except for the extraction and sorting of data, which were performed in Matlab 6.5 using the NetCDF software (http://my.unidata.ucar.edu). Likewise, matlab files (m-files) written by this author and

used for data import, pre-processing, analysis and evaluation are enclosed separately for each paper. Standard algorithms for data pre-processing and two-way data analysis were downloaded from www.models.kvl.dk and www-its.chem.uva.nl/research/pac. The N-way toolbox (51) was downloaded from www.models.kvl.dk and used for Parallel Factor Analysis.

2.1 Sample preparation and chemical analysis

A wide selection of analytical instrumental techniques can be used for the separation and analysis of petroleum hydrocarbons. Although my work focuses on fluorescence spectroscopy and GC-MS, the overall method for analysis of complex chemical mixtures can be extended to other hyphenated analytical techniques such as GC-FID and LC-MS. Fluorescence spectroscopy and GC-MS have been used for screening and more comprehensive compound-specific analysis respectively. Fluorescence spectroscopy combined with three-way decomposition of spectra (PARAFAC) was applied in Paper IV for initial screening of oil samples with respect to chemical fingerprinting. GC-MS on the other hand was used for compound-specific analysis of petroleum biomarkers respectively in Paper I, II and III and PACs in Paper I, II and V with respect to environmental forensics (Paper II and III) and for studying oil biodegradation (Paper V).

GC-MS and fluorescence spectroscopy can generate extensive amounts of data when applied to the analysis of complex chemical mixtures. Standard cleanup and peak quantification procedures involve a series of time-consuming steps, e.g., extraction, evaporation, fractionation, addition of surrogate standards, peak detection and integration and quantification based on response factors. Consequently, standard analytical procedures for oil analysis have generally been limited to either quantification of relatively few compounds in the mixture or subjective pattern matching, by e.g., visual comparison of GC-MS chromatograms or fluorescence spectra.

The purpose throughout this work has been to develop and apply analytical procedures with limited time-consumption. In the same time a consistently high data quality is assured by applying comprehensive quality assurance and quality control measures (QA/QC). Hence, reduction of the analysis time simultaneous with improved data pre-processing and data analysis allows for a more adequate, comprehensive and objective analysis of complex chemical mixtures.

Sample preparation

Sampling steps have been limited to sampling of pure oil collected on board ships, from vegetation and stones after oil spills (Paper II, III and IV), and to sacrificing experimental units (Erlenmeyer flasks) during *in vitro* experiments (Paper V). Likewise, sample preparation (i.e., extraction and cleanup) has been limited to dilution of pure oil and extraction from stones and vegetation and to liquid/liquid extraction by dichloromethane. Subsequently, water and particles have been removed by cleanup through a funnel with glasswool and sodium sulphate.

Although, the sample preparation step has been limited, the protocol stated in Figure 2 can be applied to more complicated matrices such as sediments, soils and biota. In such cases, semiautomatic extraction procedures, such as microwave extraction (52,53) and accelerated solvent extraction (54), are recommended relative to less automated and more time consuming extraction procedures such as soxhlet extraction. Accelerated solvent extraction is especially attractive since extraction of 1-24 samples can be performed overnight with limited use of solvent (e.g., 10-30 ml per sample) and sample cleanup can be included as an integrated part of the extraction procedure. Fractionation into aliphatic, aromatic and polar fractions, which is frequently used as a cleanup step for oil analysis (55,56), has been avoided throughout the present work to reduce the analysis time. However, chemical analysis of more complex sample matrices may require a fractionation step.

Semi-quantitative analysis

The general analytical method used throughout this work is based on a semi-quantitative approach. Quantitative analysis relies on either the internal or external quantification method, where the former includes addition of surrogate standards. Conversely, the semi-quantitative approach relies on frequent analyses of a laboratory reference sample. Its sample characteristics and chemical composition need to be comparable with those of the analytical samples. In Paper I, II, III and V the reference was a 1:1 mixture of a North Sea Crude oil (Brent crude oil) and a HFO from the *Baltic Carrier* oil spill. In the fluorescence screening study (Paper IV) the reference was a mixture of Brent crude oil, a light fuel oil (LFO), a HFO and a lubricating oil (Lub) prepared in such a way that the four oils contributed approximately equally to the combined fluorescence signal. Hence, the sample characteristics of the references were comparable with those of the analytical samples, and further included most of the peaks relevant for GC-MS analysis

Quality assurance and quality control (QA/QC)

The replicate references were often analysed and were used extensively for QA/QC, to calculate the analytical uncertainty, optimise the pre-processing parameters (Paper III, IV and V), for normalisation of diagnostic ratios (Paper I and II) and in the peak matching procedure of Paper I. The following QA/QC protocol based on the replicate references was used throughout the analytical GC-MS work:

- The chromatographic peak shapes were checked regularly. 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 frequently).
- Mass discrimination due to changes in the conditions of the inlet and a dirty ion source (also checked by tuning the mass spectrometer frequently).

The above QA/QC protocol has been performed daily by comparing peak shape and intensities at m/z 85 (*n*-alkanes and isoprenoids), 128 (naphthalene), 180 (methylfluorenes), 191 (terpanes), 192 (methylphenanthrenes), 198 (methyldibenzothiophenes), 217 (steranes) and

252 (five-ring PACs). Changes in the chromatographic peak shapes, mass discrimination and a significant decrease in sensitivity (more than a factor of ten) immediately led to cleaning of the ion source and prefilter, change of liner and septum or trimming of the capillary column.

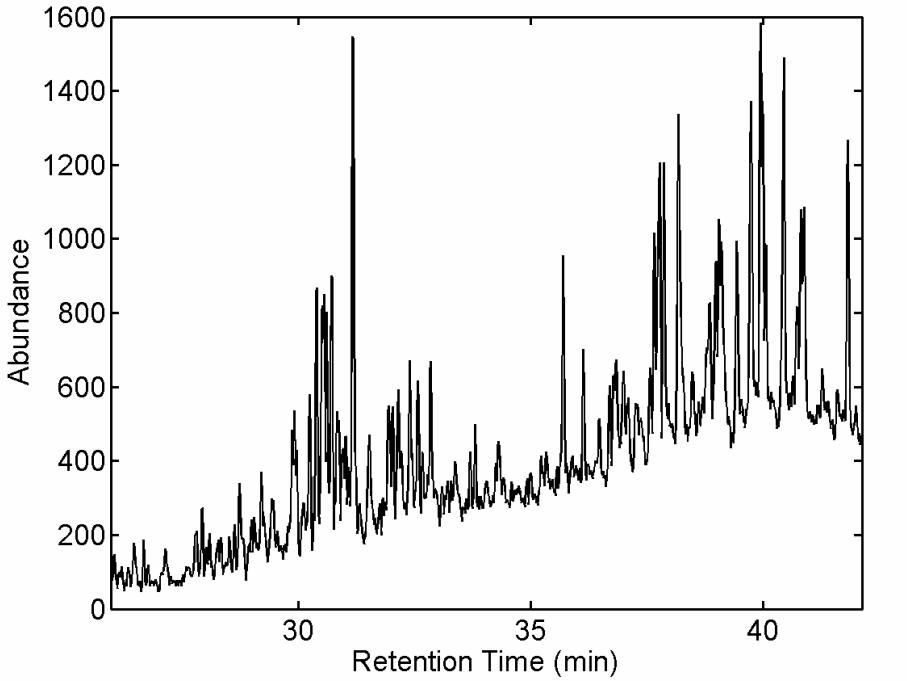


Figure 3. Chromatographic profile of m/z 217. This profile contains tricyclic steranes (eluting between 26 and 34 min) and tetracyclic steranes (eluting between 35 and 42 min).

Since oil is a "dirty" matrix however, and since no cleanup was performed prior to injection, the septum and liner (1 mm id) were changed and the ion source cleaned regularly (for every 50 injections) throughout the analytical work. This resulted in chromatographic data of consistently high quality. In fact, although more than 1000 oil samples were analysed during this work, only a few batches of samples (a batch consists of 50 - 100 samples including blanks and references) were rerun due to insufficient data quality, i.e., poor reproducibility or sensitivity, mass discrimination and chromatographic resolution.

High data quality is a prerequisite for high quality data analysis. In Paper I and II the relative analytical standard deviations (RSD_A) were consistently less than 3% for diagnostic ratios of well-resolved peaks. In paper III, IV and V the variability of replicate reference samples in score plots were much lower than the total variability within data sets. A high data quality is always important within analytical and environmental chemistry. However, during the work with time warping and PCA (Paper III and V), a high data quality was of special importance. Variations in peak shape (e.g., from symmetrical to tailing) during column deterioration affect the multivariate data analysis negatively, due to changes in intensity distribution of adjacent retention times within a peak region. Standard quantification methods based on peak quantification are less affected by these fac-

tors, since peak areas and heights are relatively independent of peak shape.

GC-MS method

In the initial phase of this work, a preliminary study was performed to decide on the compounds of interest for oil analysis with respect to environmental forensics and fate of oil spills. The selection of m/z ions with high sensitivity (base peaks in mass spectra) and selectivity for specific compound groups with common structure is based on prior work (1). Petroleum biomarkers and PACs with common structure such as tri-pentacyclic terpanes (m/z 191), steranes (m/z 217 and 218) and the homologue series of C_0-C_4-phenanthrene isomers (m/z 178, 192, 206, 220 and 234) can be measured by detecting characteristic mass fragments or the molecular masses. Hence, more than 100 m/z fragments were measured in the preliminary study and 48 mass fragments were selected and analysed in six groups of 14-15 ions using SIM in Paper II and III with data acquisition time of 1.27 scans/s.

Table 1: Complete list of mass fragments and corresponding compound groups analysed in a single GC-run.

Polycyclic aromatic hydrocarbons	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-biphenyles	154, 168, 182, 194

In Paper II and III, however, it was found that data pre-processing could be improved by increasing the sampling rate (e.g., focussing on fewer masses in each segment in the GC-MS (SIM) analysis and decreasing the dwell time. Likewise, the initial parameterisation and

peak fitting procedure applied in Paper II and described in Paper I would be affected positively, since the number of data points in each peak would increase correspondingly. In Paper II the early eluting peaks such as naphthalene (m/z 128) consisted of 8 - 10 data points, which was sometimes insufficient for determining the initial parameters for the gaussian peak fits. Furthermore, in Paper III it was concluded that an increased sampling rate would allow for more refined corrections in correlation optimised warping (COW). As a consequence the analytical procedure was changed slightly for Paper I and V such that 44 mass fragments were analysed in 8 groups of 12 ions with a data acquisition time of 2.34 scans/s. This almost doubled the number of scans per second.

Fluorescence spectroscopy

The experimental procedure used in Paper IV for fluorescence spectroscopy is based on dilution to avoid pronounced light absorption. Thus, inner filter effects and effects of quenching and energy transfer processes are reduced. The combination of high detector voltage (850 Volt), necessary to enable 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 modelling the systematic variations and leaving the noise in the residuals (see Paper IV). Excitation-emission scans (EEMs – excitation-emission matrices) were measured with excitation wavelengths ranging from 240 – 475 with 5 nm increment and emission scans from 250 – 600 nm with 2 nm increment.

The appropriate scan ranges used in Paper IV were determined in a preliminary study, demonstrating that below excitation of 240 nm the spectra were very noisy and above 600 nm the signals were negligible for the four oil types (crude oils, LFOs, HFOs and Lubs). The fluorescence EEM measurements were thus consistent with the general aim of this work in being fast and with only limited sample preparation (dilution and UV-VIS measurements). An excitation-emission scan of a HFO sample from the cargo tank of the *Baltic Carrier* is shown in Figure 4.

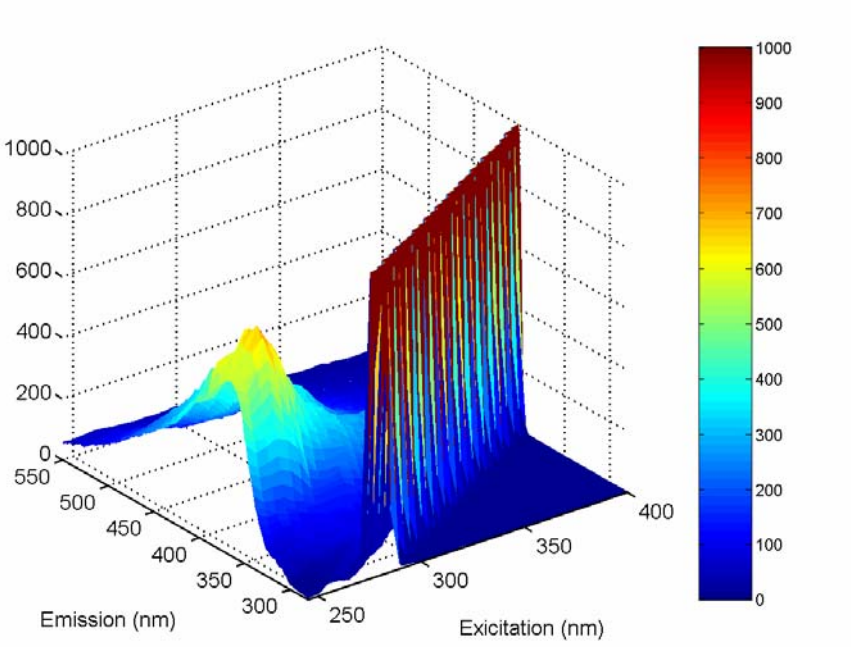


Figure 4. Fluorescence excitation-emission scans of a HFO. Modified from Paper IV.

2.2 Data pre-processing

The philosophy of the data pre-processing step is to reduce variation in the chemical data unrelated to the chemical composition such as analytical variability and concentration effects. A further requirement is that the tools are rapid and semi-automated with limited human intervention. Three types of data were treated in Paper I – V: sections of chromatograms, diagnostic ratios and fluorescence EEM spectra, and the required pre-processing are different for the three types of data.

Sections of chromatograms

A selection of procedures, including baseline removal, normalisation and chromatographic alignment, was used in Paper III and V, in order to pre-process sections of chromatograms prior to chemometric data analysis. The pre-processing procedure and the automated optimisation using replicate reference samples have been thoroughly described in the two papers and will not be described further here. The pre-processing procedure was generally effective in removing variation unrelated to the relative chemical composition. Thus, the variation of replicate references and *Baltic Carrier* spill samples were small, in the score plots of Paper III and V, compared to the total variation in the data sets.

The maximal residual misalignment after time warping was one point. The consistently high warping quality of COW on GC-MS (SIM) data is most likely due to the fact that compounds present in each of the individual chromatograms, e.g., tri- and tetracyclic steranes (Paper III) and methylated PAC homologues (Paper V) have similar physicochemical properties. Thus, they are affected in the same way by changes in the column properties such as chemical

changes in the stationary phase. The effects of baseline removal (by calculating the first derivative), normalisation and time warping are illustrated in Figure 5.

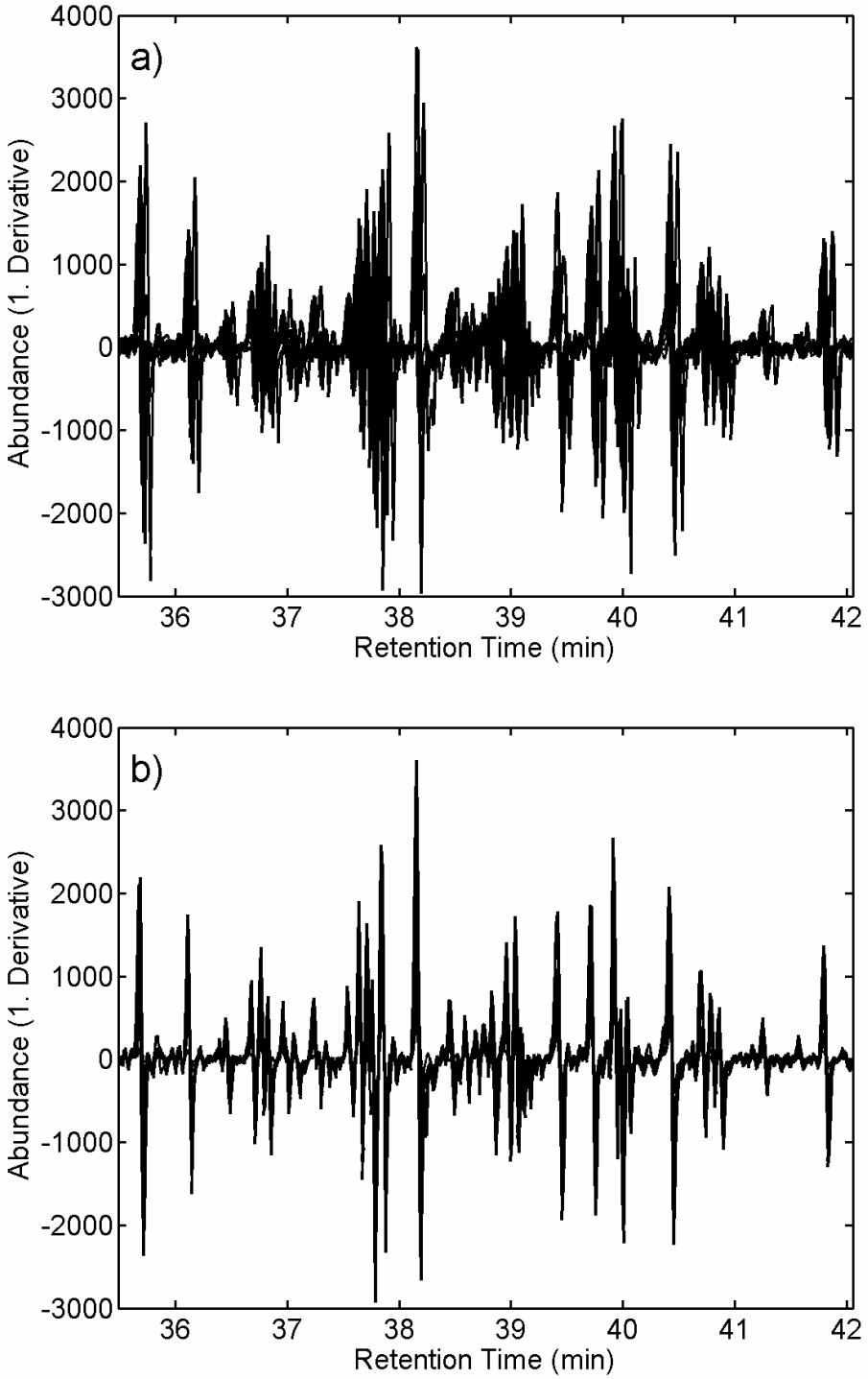


Figure 5. First derivative of a section of m/z 217 for five reference oils and five source oils: (a) before warping and (b) after warping using COW with segment length of 175 data points and a slack of 3 points. Modified from Paper III.

Conversely, chromatographic methods with less selective detectors such as GC-FID and liquid chromatography – diode array detection (LC-DAD) lead to chromatograms with more irregular retention time

shifts, since the chromatograms contain compounds with very different physicochemical properties. In situations with large shifts in the stationary phase of the column, peaks may even change elution order, and COW cannot adequately correct for such changes. This may explain why some authors (57,58) have experienced that optimal chromatographic alignment is achieved using segment lengths of the same size as individual peaks, since a high flexibility is necessary to correct for irregular retention time shifts.

COW combined with PCA can be used for analysing chromatographic data in situations where the amount of systematic information is larger than the variations caused by insufficient alignment. As an increasing number of components is extracted from the data matrix by for example PCA, the variation caused by systematic variation decreases until insufficient alignment contributes to the most pronounced variation. In Paper III, four components were extracted from the calibration set of 61 × 1231 consisting of the profile of tri- and tetracyclic steranes (m/z 217) in a set of 61 crude oils and refined petroleum products. The amount of independent information was less in Paper V and two components were extracted. The residual shifts were confounded to all subsequent components.

Diagnostic ratios

Diagnostic ratios can be used to characterise the oil composition, as is the case for pre-processed sections of chromatograms. Diagnostic ratios can be calculated either from quantitative (i.e., compound concentrations) or semi-quantitative (i.e., peak areas or heights) data. The former is most frequently used in the literature (6,48) since the uncertainties are supposed to be small and less affected by changes in gas chromatographic conditions or sensitivity changes. This is, however, only the case if the surrogate standards have similar chemical properties as the quantified compounds. Hence, the larger the diffe-rences in chemical properties are between surrogate standards and quantification standards, the less is corrected for this variability. Thus, application of a large number of standards is a requirement for sufficient normalisation of the individual compounds in oil. The concentrations are calculated from a calibration curve (a model), and the values depend on the quality (uncertainty) of the model (usually li-near). In addition, the more surrogate and quantification standards that are required, the more time consuming and expensive becomes quantitative analysis. Addition of surrogate and quantification standards to oil samples further increases the uncertainty in unpredictable ways depending on the precision of the addition process.

Calculation of diagnostic ratios from peak areas or heights, on the other hand, is less time-consuming, and the manual addition of surrogate standards and compound quantification by calibration models do not introduce uncertainty. The disadvantage of a semi-quantitative approach is, however, that the variability caused by changes in gas chromatographic conditions or in the sensitivity of the mass spectrometer is not reduced by the use of surrogate standards. This means that diagnostic ratios of compounds with quite different chemical properties may change systematically over time due to changes in the instrumental conditions not corrected for. However, this can be corrected for by normalising each ratio to the correspond-

ing ratio in the most recent analysed laboratory reference sample. Paper II demonstrates that external normalisation systematically reduced the influence of measurement errors. The use of semi-quantitative as opposed to quantitative analyses has been a matter of discussion in the scientific community. In Paper II, the analytical standard deviation of normalised diagnostic ratios was between 0.05% and 3.2% which is comparable to or smaller than those obtained by a quantitative method (48).

In software packages of mass spectrometers it is unlikely to select one set of peak identification and quantification parameters, that are optimal for hundreds of peaks with different signal-to-noise ratio, chromatographic resolution and shape. Thus, quantification of multi-component mixtures in a large number of samples is time-consuming and costly without automated chromatographic data processing. Paper I presents a semi-automated procedure for processing GC-MS (SIM) data to provide fast and reproducible transformation of chromatographic data into diagnostic ratios. The method is based on automatic peak matching, initial parameterisation, alternating background noise reduction and peak estimation using mathematical functions (Gaussian and exponential-Gaussian hybrid) with three to four parameters. The data pre-processing method has gradually evolved throughout the thesis. The crucial step is automated peak matching and alternating background noise reduction and peak estimation. These two procedures offer objective peak quantification with limited amount of human intervention. Thus, manual parameter adjustment is unnecessary for baseline distorted and incompletely separated peaks. This is not the case in processing software included in standard software packages. Frequent analysis of the reference is a requirement for the peak matching algorithm, since for each peak, the difference in shift between an analytical sample and the closest reference sample is required to be less than half the distance between neighbouring peaks. Otherwise, peaks may be wrongly assigned. The chromatographic pre-processing procedure described in Paper I has been used for pre-processing data in Paper II. In Paper V, on the other hand, Xcalibur ver.1.3 (commercial integration software) was used, since the number of peaks and samples was limited in this study. Furthermore, implementation of the pre-processing method at that time had problems in quantifying peaks close to the detection limit.

Pre-processing of fluorescence spectra

A selection of pre-processing procedures was also applied for reducing variation unrelated to the relative PAC composition in fluorescence EEMs. The effects of Raman and Raleigh scatter, which cannot be modelled by PARAFAC, was reduced by subtracting blanks (dichloromethane) and inserting missing values in the lower right triangle of EEMs. In addition to excitation-emission pairs with emission wavelength lower than excitation, diagonal lines with emission wavelengths from 0 to 10 nm higher than excitation were included, which removed data still affected by Raleigh scatter after the subtraction of blanks (Paper IV). EEMs were subsequently corrected for instrument biases by applying an excitation/emission correction, which removed small artifacts in EEMs due to variations in detector efficiency as a function of wavelength. A pre-processed EEM is shown in Figure 6.

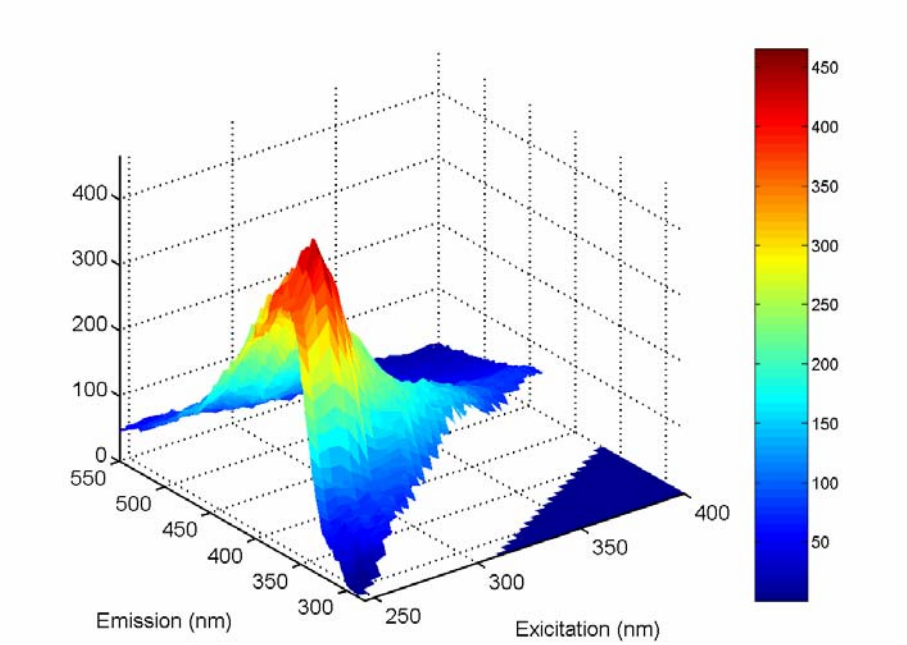


Figure 6. Pre-processed fluorescence EEM after blank subtraction, insertion of missing values and a small triangle of zeros and excitation/emission correction. Modified from Paper IV.

2.3 Data analysis

Multivariate statistical methods such as PCA and PARAFAC work by approximating the "best" model of data (often in a least-squares sense). Hence, a model can be a more or less appropriate model of data, and the parts not described by the model are left in the residuals. The size of these residuals can vary from sample to sample and between variables leading to a more or less adequate description of individual samples and variables (e.g., samples and variables with high residuals are poorly described by the model). Consequently, if a multivariate model does not describe data sufficiently, the interpretation can lead to inadequate and even misleading conclusions in for example environmental forensic investigations.

One advantage of multivariate compared to univariate statistical methods is the ease by which relations between multiple samples and variables (e.g., compound concentrations or diagnostic ratios) can be resolved and visualised using score and loading plots.

Another advantage of multivariate methods is noise reduction. Noise reduction is obtained when more than one variable describes the same phenomenon (i.e., interrelated variables). Examples of interrelated diagnostic ratios are heptadecane/pristane and octadecane/phytane, which describe the same phenomenon, namely biodegradation. The ratios are highly selective because the physicochemical properties of their constituents (e.g., boiling point and water solubility) are almost identical, but the branched alkanes are less susceptible to microbial degradation compared to the linear alkanes. Likewise, if

several diagnostic ratios separate oil samples due to the same underlying phenomenon (e.g., depositional environment, thermal maturity and in-reservoir degradation) the distinction becomes less and less affected by noise. Reduction of the uncertainties due to interrelated variables and modelling of these variations corresponds to taking the mean of replicate samples. The more replicates that are used to calculate the mean, the more certain it becomes ($\bar{x}$ approaches μ, which is the true mean). For example, in PCA the information in many correlated variables are summarised into few principal components that are weighted sums of the original ones, where the weights are the loadings. The approach of including se-veral variables in the data analysis that in essence describes the same phenomenon opposes univariate data analysis, where a few descriptive ratios are selected from a suite of ratios potentially describing the same phenomenon.

Although data are pre-processed prior to the multivariate statistical data analysis (Figure 2) in order to reduce information unrelated to the chemical composition, some variables (chromatographic retention times, diagnostic ratios or excitation-emission pairs) are more informative than others are. Two approaches have been used in Paper II, III and V to either deselect variables prior to chemometric data analysis or to fit the principal component model according to a weighted least squares (WLS) criterion. In Paper II the inverse of the relative sampling standard deviation (RSD_S) was used as weights, whereas in Paper III and V the inverse of the RSD_A was used. The resolution power has been used throughout the work as a criterion for optimising the pre-processing and data analysis (including variable selection and WLS-PCA). The resolution power has been defined as the ability of the PCA model to distinguish dissimilar samples. WLS-PCA gave generally 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 data set (e.g., analytical uncertainty). If this is not the case, WLS-PCA may result in a model with lower resolution power than PCA without variable selection and to a biased model.

2.4 Data evaluation

Visual inspection of score and loading plots

Data were evaluated based on visual inspection of score and loading plots (Paper III, IV and V), since relations between multiple samples and variables can easily be resolved and visualised. Score plots map the main relationships between samples based on the original variables. In Figure 7 the second principal component (PC2) is plotted versus the fourth principal component (PC4), using data from Paper III. Oil samples closely located in score plots have similar chemical composition based on the original variables (m/z 217 chromatogram). Conversely, oil samples located far apart in the score plots have dissimilar chemical composition, and this dissimilarity increases as the distance increases. Note that principal components are ordered according to their explained variance in PCA, with the first compo-

nent describing most of the variation and subsequent components (e.g., PC2, PC3, PC4 etc) explaining a decreasing amount of variation. Hence, differences along the first components are more significant than along higher order components.

Although the higher order components describe a lesser percentage of the total variance they are not necessarily of lesser importance for the separation of dissimilar oil samples. Especially for large data sets, an important separation between similar samples is often found in higher order components. In Paper III for example, the suspected source samples Oseberg South East (SE) and Oseberg Field Centre (FC) cluster along PC1, PC2 and PC3 but are well separated along PC4 which is a minor component describing only 6.27% of the total variation in the calibration data set (see Figure 7).

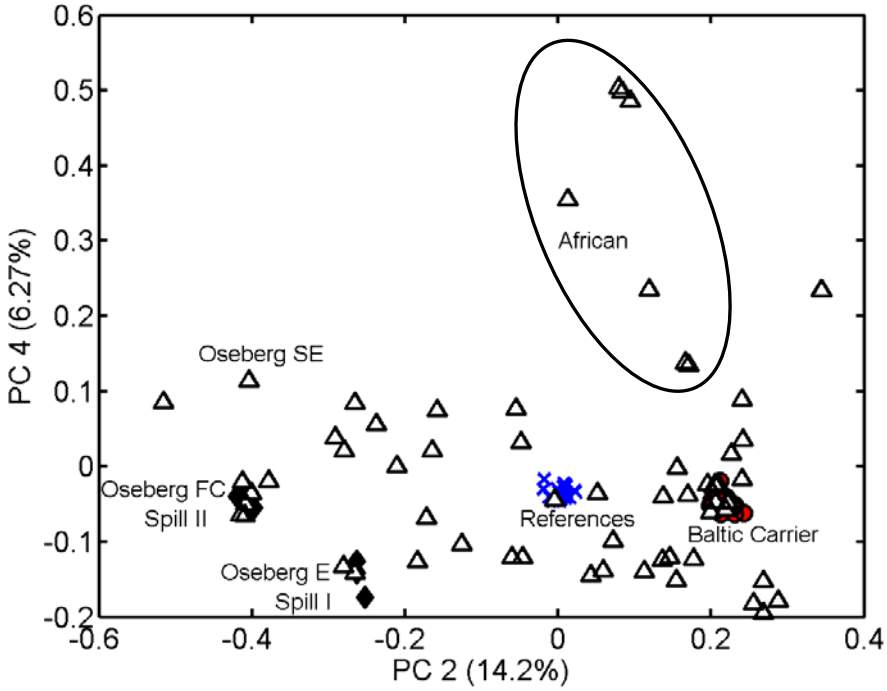


Figure 7. Score plot (PV2 vs PC4) using WLS-PCA on sections of pre-processed GC-MS (SIM) chromatograms of m/z 217. The PCA model was calculated from the calibration set (61 × 1231), whereas the reference set (18 × 1231) and test set (22 × 1231) were calculated by projecting the data onto the loadings.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

In Paper II local PCA models (i.e., modelling a subset of closely located samples) were used to focus the data analysis on separating related samples with similar biomarker and PAC 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 order PCs. The latter are more affected by noise since they describe only a small percentage of the total variation in the data set.

In Paper IV the normalised PARAFAC scores were used to characterise and match oil samples based on their relative PAC composition. The four oil types (crude oils, HFOs, LFOs and Lubs) could be distinguished in the score plots (except for some overlap of HFOs and crude oils). The oil characterisation showed that crude oils have the most even distribution of factors and a large variability in the content of high molecular weight PACs. In contrast, LFOs and Lubs have a high relative content of low molecular weight PACs, whereas HFOs have a high relative content of high molecular weight compounds.

In Paper V the score plot of PC1 vs PC2 was used to evaluate the effects of biodegradation on the relative amounts of isomers of methylfluorenes, methylphenanthrenes and methyldibenzothiophenes. It was concluded that the three mixtures of bacterial strains (R, U and M) affected the isomer distribution in different ways. Hence, the study illustrated that one should be careful about deducing the preferential degradation of PAC homologues in the natural environment from observations made during *in vitro* experiments using a simple mixture of bacterial strains. A selection of diagnostic ratios and sections of pre-processed chromatograms were used in this study.

Similarity indices

The objectivity of the matching process using time warping and PCA for environmental forensics can be improved by comparing samples using similarity indices (Paper IV) or statistical tests (Paper II). In Paper IV similarity of oil samples were calculated using the correlation coefficient based on the similarity of normalised 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 scores could also have been used in Paper III for objective spill/source matching of pre-processed chromatographic sections of biomarker hydrocarbons.

Statistical tests

An even more objective method for matching oil samples was applied in Paper II. The method consisted 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 PCA model (i.e., the retained PCs) was tested independently accepting a 5% error level, $\alpha = 0.05$. If the inequality was false in at least one of these tests, the overall H_0 was rejected, and the tested source oil was 'beyond reasonable doubt' not the source of the spill. Since the method includes multiple comparisons, the risk for an overall type one error (i.e., that H_0 is rejected when it is true) increases with the number of comparisons. In Paper II the Bonferroni correction of the α-value was applied to compensate for this.

Interpretation of chemical fingerprinting results

Interpretation of the results from chemical fingerprinting facilitates comparisons of source oils and spill samples. Variables (e.g., retention times or diagnostic ratios) contributing most to a PC are associated with large negative or positive coefficients in the corresponding loadings. Panels a and b in Figure 8 show part of the cumulative sum of the PC2 and PC4 loadings, respectively for the chemical finger-

printing study in Paper III based on pre-processed chromatograms of petroleum biomarkers.

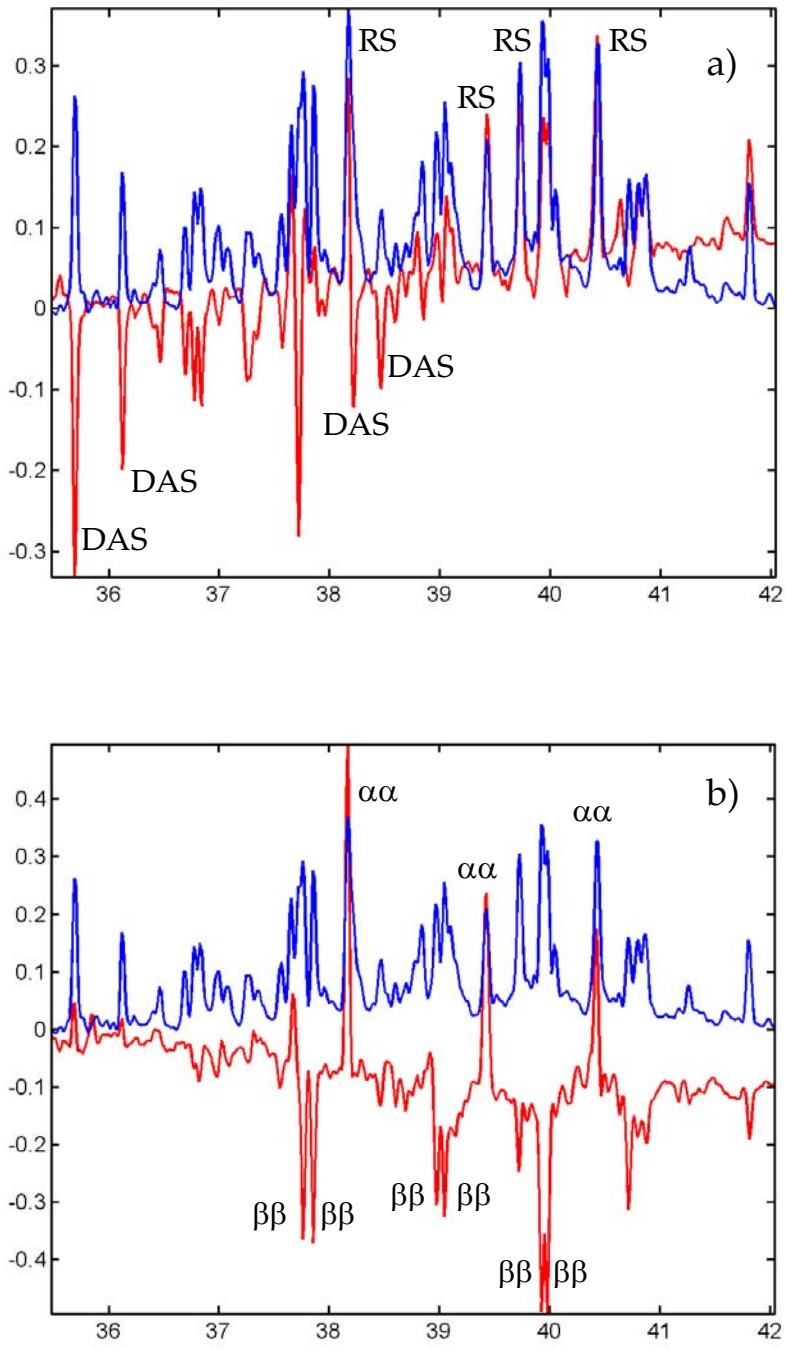


Figure 8. Integrated mean-chromatogram (blue) and integrated loadings (red) for WLS-PCA. Panels a and b show the PC2 and PC4 loading, respectively. Selected rearranged steranes (RS), diasteranes (DAS), αα-steranes (αα) and ββ-steranes (ββ) are marked in the plot.

The PC2 loadings (Figure 8a) are negative for diasteranes (DAS) and positive for rearranged steranes (RS). Ratios of the type DAS/RS are commonly used to distinguish oil originating from source rocks with different clay content (1). Low DAS/RS indicates anoxic clay-poor carbonate source rock, whereas high DAS/RS indicates source rocks abundant in clays. Hence, PC2 can be interpreted as a source parame-

ter where oil samples with high PC2 (e.g., the *Baltic Carrier* oil) are derived from a source rock containing less clay than oil samples with low PC2 (e.g., North Sea crude oils). The PC4 loadings (Figure 8b) are negative for ββ-isomers of C_{27} to C_{29}-rearranged steranes and positive for αα-isomers. The ratio of C_{29}-rearranged steranes (ββ / (ββ+αα)) is a highly specific parameter for maturity, and appears to be independent on source organic matter input (1). The ββ isomers have a higher thermal stability than αα isomers; thus the above ratio increases with thermal maturity. African crude oils have positive scores for this component and, thus, have low relative maturity. Furthermore, PC1 and PC3 (not shown) describe respectively the boiling point range and carbon number distribution of sterols in the organic matter of the source rock.

Interpretation of biodegradation results

In fate studies the loading plot is especially important, because it shows which original variables (e.g., retention times or diagnostic ratios) are responsible for the directions, changes and groupings observed in the corresponding score plot. The loading plot was used in Paper V to interpret the changes observed in the corresponding score plot of PC1 vs PC2. More specifically, it was used to determine the general degradation order of PAC isomers within homologue series and the differences between the effects of the three bacterial inocula on oil composition. The PC1 scores of oil samples generally increased with increasing biodegradation and, thus, isomers with highest negative loadings compared to the unweathered oil were most susceptible to microbial degradation and so forth. Using this simple evaluation technique the degradation order was determined from the three groups of PACs as: 2MP = 3MP > 1MP > 49MP for methylphenanthrenes; 3MF > 1MF > 2MF > 4MF for methylfluorenes and 2MD > 1MD > 4MD > 3MD for methyldibenzothiophenes. Likewise, evaluation of the PC2 scores with the corresponding PC2 loadings indicated that the M-samples had a less preferential degradation of 1MF (most positive loadings) over 2MF (most negative loadings) and 1MD over ∑23MD. Furthermore, a more preferential degradation of 1MP (most negative) over 49MP (most positive) was found compared to the U- and R-samples.

In Paper IV, the PARAFAC factors were interpreted by comparing excitation and emission loadings with EEMs of selected PACs and the fluorescence characteristics for a broad range of PACs. Since PARAFAC factors can be uniquely determined up to trivial permutation and scaling (59,60) the factors describe underlying chemical spectra. The comparisons indicated that each of the factors could be related to a mixture of PACs with similar fluorescence characteristics: A mixture of naphthalenes and dibenzothiophenes, fluorenes, phenanthrenes, chrysenes and five-ring PACs.

3 Conclusion

Rapid, reliable and objective tools are a requirement for the characterisation of complex chemical mixtures. This thesis describes the development of such tools for assessing the identity (i.e., in terms of environmental forensics) and fate (i.e., weathering) of petroleum hydrocarbon mixtures. An overall integrated methodology comprising four steps was developed and applied throughout the thesis: Chemical analysis, data pre-processing, data analysis using chemometrics and data evaluation.

The chemical analyses were based on fluorescence spectroscopy and GC-MS (SIM). The fluorescence spectroscopy method facilitated a fast (less than 10 min. per sample) screening and characterisation of oil samples based on their main composition of PACs. The technique, thus provides a fast and alternative, yet complementary (revealing PAC 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 semi-quantitative approach including frequent analysis of a reference sample. With respect to environmental forensics and fate studies, the approach based on 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 for external normalisation. The semi-quantitative approach based on GC-MS (SIM) constituted as such an important part of the integrated methodology for analysis of complex chemical mixtures.

The pre-processing tools comprised: semi-automated peak identification and quantification (Paper I and II), analysis of sections of chromatograms by baseline removal, time warping and normalisation (Paper III and V) and automated pre-processing of fluorescence EEMs (Paper IV). Time warping combined with PCA is a rapid and objective approach for analysing complex chemical mixtures compared to peak identification and quantification. The results in this thesis 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 data set and the variability due to insufficient alignment is the important criterion for, whether the method provides an appropriate pre-preprocessing tool or not, in the analysis of complex chemical mixtures.

The use of multivariate statistical methods such as PCA and PARAFAC are the cornerstone of the integrated methodology. The methods enable the analysis and assessment of large data sets by extracting a number of principal components or factors that describe the promi-

nent 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 data set (e.g., analytical uncertainty). If this is not the case, the subjectivity of the data analysis increases, and the model becomes increasingly biased.

Rapid and objective analyses of complex mixtures were attained using three evaluation techniques. Although excellent for monitoring and assessing the fate of complex chemical 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 were applied in this thesis (Paper II and IV). The analytical and sampling uncertainties were used in Paper II to test the null-hypothesis and determine the source of spill samples. Thus, the conclusions are less dependent on data variations and subjective decisions.

Each of the four steps in the integrated methodology contributes to a rapid, objective and comprehensive analysis of complex chemical mixtures. Standard methods can be employed in one or several of the steps, but that would lead to a reduction of the appropriateness of the integrate methodology for analysing complex chemical mixtures.

The methods developed within this thesis can be employed in routine investigations, and they represent major improvements compared to standard methods for analysis of complex chemical mixtures. The limited human intervention required, and the extended amounts of chemical information that can be generated, analysed and evaluated are the major and obvious strengths of this integrated methodology. More specifically, the methods described in papers II, III and IV 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. 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 pre-screening oil samples (Paper IV), while GC-MS (SIM) combined with fast and objective pre-processing, data analysis and data evaluation (Paper II and III) can be used for compound-specific fingerprinting.

Although this thesis has focused on petroleum hydrocarbon mixtures (crude oil and petroleum products), the integrated methodology and the individual methods have a broader applicability. Data from other analytical instruments such as LC-MS and data of other complex mixtures than hydrocarbons can be pre-processed, analysed and evaluated. Furthermore, the methods can be used for monitoring, environmental forensics and risk assessment of other contaminant mixtures in environmental samples. In fact, a combination of compound concentrations, diagnostic ratios and PCA has been used in the associated papers A, B and C to evaluate the pattern of haloge-nated compounds in marine biota (e.g., fish, mussels, black guillemots and other species) from Greenland.

4 Perspectives

Risk assessment is often limited to one chemical or one group of closely related compounds at a time. Likewise, environmental monitoring typically employs specific and tiered analyses of relatively few compounds. The importance of assessing complex chemical mixtures will probably get a higher priority in future research than has been the case so far. The methods described in papers I – V are in that respect major improvements compared to standard methods. Yet a series of further developments may improve their use for the characterisation of complex chemical mixtures.

The ability to distinguish samples from different sources in environmental forensics can be enhanced by including more compounds in the analysis (i.e., by adding more characteristic MS fragment ions), as was done in Paper V for assessing the microbial degradation of oil. Combining several GC-MS (SIM) chromatograms (e.g., terpanes, steranes and PAC homologue series) will most likely improve the ability of the PCA model to distinguish between dissimilar oil samples. Furthermore, correlation indices can be used for comparing sample scores from PCA as was accomplished with PARAFAC factors in Paper IV.

The developed methods were used in Paper V to study the environmental fate of petroleum hydrocarbons. The methods can be extended to the evaluation of physical and chemical weathering processes by adding information from other fragment ions. The relative changes in the homologue series of *n*-alkanes can for example be used as an indication of evaporative effects (depend on boiling points variations), while the relative ratios of *n*-alkanes to isoprenoids describe the initial microbial degradation.

With small adjustments these methods will also be applicable to other complex chemical mixtures. The requirements for time warping include for example that SIM chromatograms are available of compounds with similar physiochemical properties, as that will ensure regular retention time shifts. Furthermore, the chromatograms should contain complex peak patterns to maximise information compared with the inevitable residual retention time shifts in data after warping. If these requirements are fulfilled, the time warping approach can be directly applied. This is currently being studied for polychlorinated terphenyls in environmental samples, primarily sewage sludge. The purpose of this study is to characterise the polychlorinated terphenyl patterns in the environmental samples in relation to three technical mixtures used in industry. If the compounds of interest have less complex patterns, the variation caused by insufficient alignment may become the most pronounced variation in the data. In such a case, PCA will be unable to describe the systematic patterns related to the chemical composition. This obstacle may be overcome by refining the time warping approach, e.g., by using longer segment length and data interpolation prior to warping.

Another future prospect is to retain the concentration effects in data. This can be achieved by normalising to one or several surrogate standards, (which corresponds to the internal quantification method) or by using the normalisation factor of the closest analysed reference for normalisation (corresponds to the external quantification method). Monitoring of concentrations of contaminants will furthermore facilitate the risk assessment of complex mixtures.

5 Literature cited

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

(2) Pharr,D.Y., Mckenzie,J.K., and Hickman,A.B., *Ground Water*, 30 (**1992**) 484-489.

(3) Siegel,J.A., Fisher,J., Gilna,C., Spadafora,A., and Krupp,D., *J.Forensic Sci.*, 30 (**1985**) 741-759.

(4) Wang,Z.D. and Fingas,M., *Energ.Source*, 25 (**2003**) 491-508.

(5) Daling,P.S., Faksness,L.G., Hansen,A.B., and Stout,S.A., *Environmental Forensics*, 3 (**2002**) 263-278.

(6) Wang,Z.D., Fingas,M., and Page,D.S., *J.Chromatogr.A*, 843 (**1999**) 369-411.

(7) Frysinger,G.S. and Gaines,R.B., *Journal of Separation Science*, 24 (**2001**) 87-96.

(8) Frysinger,G.S., Gaines,R.B., and Reddy,C.M., *Environmental Forensics*, 3 (**2002**) 27-34.

(9) Mansuy,L., Philp,R.P., and Allen,J., *Environ.Sci.Technol*, 31 (**1997**) 3417-3425.

(10) Rogers,K.M. and Savard,M.M., *Org.Geochem*, 30 (**1999**) 1559-1569.

(11) Munoz,D., Doumenq,P., Guiliano,M., Jacquot,F., Scherrer,P., and Mille,G., *Talanta*, 45 (**1997**) 1-12.

(12) Wang,Z.D. and Fingas,M., *J.Microcolumn.Sep*, 7 (**1995**) 617-639.

(13) Mille,G., Munoz,D., Jacquot,F., Rivet,L., and Bertrand,J.C., *Estuar.Coast.Shelf.S*, 47 (**1998**) 547-559.

(14) Prince,R.C., Owens,E.H., and Sergy,G.A., *Mar.Pollut.Bull*, 44 (**2002**) 1236-1242.

(15) Sauer,T.C., Brown,J.S., Boehm,P.D., Aurand,D.V., Michel,J., and Hayes,M.O., *Mar.Pollut.Bull*, 27 (**1993**) 117-134.

(16) Wang,Z., Fingas,M., Blenkinsopp,H., Sergy,G., Landriault,M., Sigouin,L., and Lambert,P., *Environ.Sci.Technol*, 32 (**1998**) 2222-2232.

(17) Leblond,J.D., Schultz,T.W., and Sayler,G.S., *Chemosphere*, 42 (**2001**) 333-343.

(18) Rowland,S.J., Alexander,R., Kagi,R.I., Jones,D.M., and Douglas,A.G., *Org.Geochem*, 9 (**1986**) 153-161.

(19) Thouand,G., Bauda,P., Oudot,J., Kirsch,G., Sutton,C., and Vidalie,J.F., *Can.J.Microbiol*, 45 (**1999**) 106-115.

(20) Wang,Z.D., Fingas,M., Blenkinsopp,S., Sergy,G., Landriault,M., Sigouin,L., Foght,J., Semple,K., and Westlake,D.W.S., *J.Chromatogr.A*, 809 (**1998**) 89-107.

(21) Cerniglia,C.E., *Biodegradation*, 3 (**1992**) 351-368.

(22) Douglas,G.S., Bence,A.E., Prince,R.C., McMillen,S.J., and Butler,E.L., *Environ.Sci.Technol*, 30 (**1996**) 2332-2339.

(23) Dutta,T.K. and Harayama,S., *Environ.Sci.Technol*, 34 (**2000**) 1500-1505.

(24) Oudot,J., Merlin,F.X., and Pinvidic,P., *Mar.Environ.Res*, 45 (**1998**) 113-125.

(25) Bragg,J.R., Prince,R.C., Harner,E.J., and Atlas,R.M., *Nature*, 368 (**1994**) 413-418.

(26) E.L.Butler, G.S.Douglas, and W.G.Steinhauer, On-Site Bioreclamation, 2001, p. 515-521.

(27) Prince,R.C., Elmendorf,D.L., Lute,J.R., Hsu,C.S., Halth,C.E., Senius,J.D., Dechert,G.J., Douglas,G.S., and Butler,E.L., *Environ.Sci.Technol*, 28 (**1994**) 142-145.

(28) Venosa,D., Suidan,M.T., King,D., and Wrenn,B.A., *J.Ind.Microbiol.Biot*, 18 (**1997**) 131-139.

(29) Sasaki,T., Maki,H., Ishihara,M., and Harayama,S., *Environ.Sci.Technol*, 32 (**1998**) 3618-3621.

(30) Budzinski,H., Raymond,N., Nadalig,T., Gilewicz,M., Garrigues,P., Bertrand,J.C., and Caumette,P., *Org.Geochem*, 28 (**1998**) 337-348.

(31) Kennicutt,M.C., *Oil.Chem.Pollut*, 4 (**1988**) 89-112.

(32) Christensen,L.B. and Larsen,T.H., *Ground.Water.Monit.R*, 13 (**1993**) 142-149.

(33) Volkman,J.K., *Org.Geochem*, 6 (**1984**) 619-632.

(34) Wang,Z. and Fingas,M., *Environ.Sci.Technol*, 29 (**1995**) 2842-2849.

(35) Jacquot,F., Guiliano,M., Doumenq,P., Munoz,D., and Mille,G., *Chemosphere*, 33 (**1996**) 671-681.

(36) Barakat,A.O., Mostafa,A.R., Rullkotter,J., and Hegazi,A.R., *Mar.Pollut.Bull*, 38 (**1999**) 535-544.

(37) Barakat,A.O., Mostafa,A.R., Qian,Y.R., and Kennicutt,M.C., *Spill.Sci.Technol.B*, 7 (**2002**) 229-239.

(38) Telnaes,N. and Cooper,B.S., *Mar.Petrol.Geol*, 8 (**1991**) 302-310.

(39) Chakhmakhchev,A., Sampei,Y., and Suzuki,N., *Org.Geochem*, 22 (**1994**) 311-322.

(40) Curiale,J.A., *Aapg.Bull*, 75 (**1991**) 560.

(41) Telnaes,N. and Dahl,B., *Org.Geochem*, 10 (**1986**) 425-432.

(42) Øygard,K., Grahl-Nielsen,O., and Ulvøen,S., *Org.Geochem*, 6 (**1984**) 561-567.

(43) Aboul-Kassim,T.A.T. and Simoneit,B.R.T., *Environ.Sci.Technol*, 29 (**1995**) 2473-2483.

(44) Aboul-Kassim,T.A.T. and Simoneit,B.R.T., *Mar.Pollut.Bull*,, 30 (**1995**) 63-73.

(45) Burns,W.A., Mankiewicz,P.J., Bence,A.E., Page,D.S., and Parker,K.R., *Environ.Toxicol.Chem*, 16 (**1997**) 1119-1131.

(46) Lavine,B.K., Vesanen,A., Brzozowski,D.M., and Mayfield,H.T., *Anal.Lett*, 34 (**2001**) 281-293.

(47) Mudge,S.M., *Environ.Sci.Technol*, 36 (**2002**) 2354-2360.

(48) Stout,S.A., Uhler,A.D., and McCarthy,K.J., *Environmental Forensics*, 2 (**2001**) 87-98.

(49) Christensen,J.H., *Polycycl.Aromat.Comp*, 22 (**2002**) 703-714.

(50) Simpson,C.D., Harrington,C.F., and Cullen,W.R., *Environ.Sci.Technol*, 32 (**1998**) 3266-3272.

(51) Andersson,C.A. and Bro,R., *Chemometr.Intell.Lab.*, 52 (**2000**) 1-4.

(52) Shu,Y.Y., Lao,R.C., Chiu,C.H., and Turle,R., *Chemosphere*, 41 (**2000**) 1709-1716.

(53) Jassie,L., *International laboratory news*, (**1995**) 18.

(54) Bandh,C., Bjorklund,E., Mathiasson,L., NAF,C., and Zebuhr,Y., *Environ.Sci.Technol*, 34 (**2000**) 4995-5000.

(55) Wang,Z.D., Fingas,M., and Li,K., *J.Chromatogr.Sci*, 32 (**1994**) 367-382.

(56) Wang,Z.D., Fingas,M., and Li,K., *J.Chromatogr.Sci*, 32 (**1994**) 361-366.

(57) Nielsen,N.P.V., Carstensen,J.M., and Smedsgaard,J., *J.Chromatogr.A*, 805 (**1998**) 17-35.

(58) Tomasi G., van den Berg,F., and Andersson,C., *J.Chemometr.*, 18 (**2004**) 231-241.

(59) Bro,R., *Chemometr.Intell.Lab.*, 38 (**1997**) 149-171.

(60) Riu,J. and Bro,R., *Chemometr.Intell.Lab.*, 65 (**2003**) 35-49.

Paper I

Chromatographic Preprocessing of GC-MS data for Analysis of Complex Chemical Mixtures

Christensen JH, Mortensen J, Hansen AB and Andersen O

Journal of Chromatography A, 2005, 1062, 113-125

Paper II

Integrated Methodology for Forensic Oil Spill Identification

Christensen JH, Hansen AB, Tomasi G, Mortensen J and Andersen O

Environmental Science & Technology, 2004, 38 (10), 2912-2918

Paper III

Chemical Fingerprinting of Petroleum Biomarkers using Time Warping and PCA

Christensen JH, Tomasi G and Hansen AB

Environmental Science and Technology, 2005, 39 (1), 255-260

Paper IV

Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis

Christensen JH, Hansen AB, Mortensen J and Andersen O

Analytical Chemistry, 2005, 77 (7), 2210-2217

Paper V

Multivariate Statistical Methods for Evaluating Biodegradation of Oil in the Environment

Christensen JH, Hansen AB, Karlson U, Mortensen J and Andersen O

Journal of Chromatography A
(Web Release Date: 1. August 2005)

Associated Paper A

Persistent halogenated compounds in black guillemots (*Cepphus grylle*) from Greenland – levels, compound patterns and spatial trends

Vorkamp K, Christensen JH, Glasius M and Riget FF

Marine Pollution Bulletin, 2004, 48, 111-121

Associated Paper B

Polybrominated diphenyl ethers and organochlorine compounds in biota from the marine environment of East Greenland

Vorkamp K, Christensen JH and Riget FF

Science of the Total Environment, 2004, 331, 143-155

Associated Paper C

Halogenated organic contaminants in marine fish and mussels from southern Greenland – pilot study on relations to trophic levels and local sources

Glasius M, Christensen JH, Platz J and Vorkamp K

Journal of Environmental Monitoring, 2005, 7, 127-131

Chemometrics as a tool to analyse complex chemical mixtures

National Environmental Research Institute
Ministry of the Environment

ISBN 87-7772-860-2