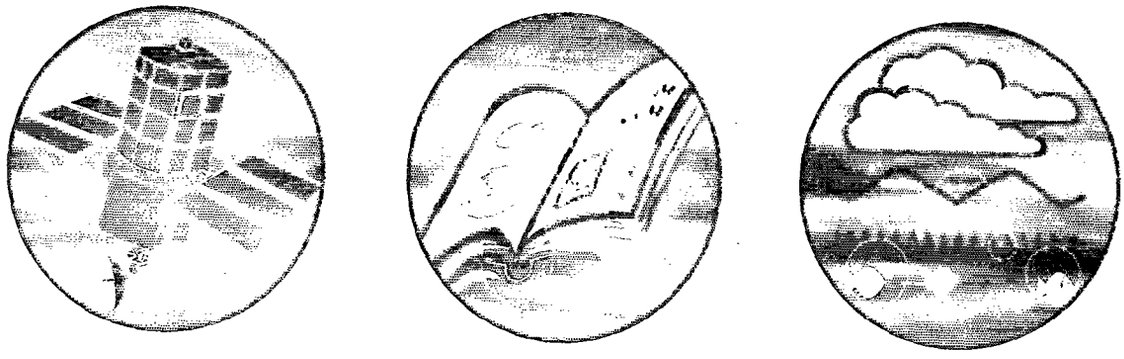
Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water



Research and Development

Technical Report
E72

ENVIRONMENT AGENCY

Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water

R&D Technical Report E72

W J Walley, P W J Robotham and M A O'Connor

Research Contractor:

School of Computing, Staffordshire University

Further copies of this report are available from:
Environment Agency R&D Dissemination Centre, c/o
WRc, Frankland Road, Swindon, Wilts SN5 8YF

tel: 01793-865000 fax: 01793-514562 e-mail: publications@wrcplc.co.uk

Publishing Organisation:
Environment Agency
Rio House
Waterside Drive
Aztec West
Almondsbury
Bristol BS32 4UD

Tel: 01454 624400 Fax: 01454 624409

ISBN: 1 85705 0304

Dissemination status
Internal: Released to Regions
External: Released to Public Domain

Statement of use
This report presents the findings of feasibility study into how computer-based methods of pattern recognition and comparison might be used to identify the origin of oil leaks from patterns in their gas chromatographs. It is intended for use by the Agency's staff and others intereste in the identifying of the source river pollution by refined oils.

Research contractor
This document was produced under R&D Project E1-050 by:

School of Computing
Staffordshire University
The Octagon
Beaconside
Stafford
ST18 0AD

Tel: 01785 353510

Environment Agency Project Manager
The Environment Agency's Project Manager for R&D Project E1-050 was:
David Britnell, Thames Region

CONTENTS

TABLES

FIGURES

APPENDICES

GLOSSARY

The following list gives brief definitions of the technical terms and acronyms used throughout this report.

ANN or NN	(Artificial) Neural Network - a computer model based on an architecture and function similar to that of the brain.
API	American Petroleum Institute.
ART-2a	Version 2a of a specific type of neural network, based on Adaptive Resonance Theory.
ASTM	American Society for Testing and Materials.
Backpropagation	An algorithm that is used to train some types of neural network.
DCM	Dichloromethane.
EPA	Environmental Protection Agency (of the USA).
Euclidean distance	The distance between two points in n-dimensional space, defined as: $D = \sqrt{\sum_{i=1}^{n} d_i^2}$ where d_i is the distance between the points w.r.t. the i th dimension.
FID	Flame ionisation detection.
FL	Fluorescence (spectrogram/spectroscopy).
GC	Gas chromatogram/chromatography.
HCA	Hierarchical cluster analysis.
HPLC	High performance liquid chromatogram/chromatography.
IR	Infrared (spectrogram/spectroscopy).
KNN	k th nearest neighbour - a statistical method of classification.
TLC	Thin layer chromatography.
MLP	Multi-layered perceptron - a form of neural network.
MLR	Multiple linear regression.
MS	Mass spectroscopy.
PCA	Principal component analysis.
PCR	Principal component regression.
PLS	Partial least squares.
Pr/Ph	Ratio of pristane to phytane.
SIMCA	Soft independent modelling of class analogy - a statistical method of classification.
SOM or SOFM	Self-organised (feature) map - a form of neural network.
UCM	Unresolved complex mixture.
UV	Ultraviolet (spectrogram/spectroscopy).

EXECUTIVE SUMMARY

This Technical Report presents the findings of National R&D Project E1-050 "The Application of Neural Networks to Pattern Recognition and Comparison - A Feasibility Study". Although the project's title specifically referred to neural networks and made no mention oil spills, its principal aim was clearly to determine the feasibility of using pattern recognition techniques to identify the source of an oil spill using chromatographic / spectrographic data. It was agreed at an early stage that the scope of the study be expanded to include statistical methods of pattern recognition, since it was found that far more research work has been done in this field using these methods than using neural networks.

A comprehensive literature search was carried out to determine the state-of-the-art and the extent to which currently available software might meet the Agency's needs. Several key researchers and consultants world-wide were contacted to seek their opinions and advice. In addition, a desk-top study of the basic nature of the pattern recognition task posed by this specific problem was carried out. The conclusions drawn from these initial studies were that:

- much work has been done using statistical methods of pattern recognition to identify the source of a crude oil spill at sea, and that a proven computer package (EUROCRUDE) is now available for this purpose;
- there are presently no computer packages based on neural networks that would be suitable for use by Agency staff for routine chemical fingerprinting of oil samples;
- the problem of identifying the source of a refined oil spill on an inland water is sufficiently different from that posed by a crude oil spill at sea as to require a different solution; and
- the current state-of-the-art makes it feasible to develop a source identification package for oil spills on inland waters using computer-based pattern recognition techniques.

A preliminary investigation, based on data supplied by the Agency, was carried out to test the viability of a potential solution to the source identification problem. The proposed solution involves a two stage process: firstly, to determine the type of refined oil responsible for the spill by comparing the fingerprints of the field samples with those of a range of standard refined oils; and secondly, to identify the source of the spill by comparing specific features of the fingerprints of the field samples with the corresponding features in the fingerprints of samples taken from potential sources. The viability of the first stage was not in doubt, since there was ample evidence in the literature to show that it can be done. However, the proposed second stage involved a novel approach, so a study was carried out to test its viability using data from a specific case study. This study was based on a visual pattern recognition exercise in which 20 volunteers attempted to match patterns in localised sections of the GC traces of field samples with those in the corresponding sections of samples derived from several potential sources. The high degree of success achieved by the participants in identifying the true source clearly indicated that an algorithm based on the matching of localised features could provide the basis for an effective computer-based source identification system.

Recommendations have been made for a phased programme of research to develop a comprehensive source identification system that would improve the reliability and defensibility of evidence present in court.

Keywords: river, pollution, liability, oil spills, chromatography, fingerprinting, pattern recognition, neural networks, chemometrics.

1. INTRODUCTION

1.1 Background to the Study

A common type of river pollution dealt with by the Environment Agency involves refined oils that have been spilled from domestic or industrial sources. The investigation of this type of pollution requires that the pollutant be characterised and compared with oils from possible sources. This is done by generating gas chromatograms or infrared spectra of both the pollutant and the oils from the potential sources, and then comparing the patterns generated by each. Where these comparisons indicates a particular source they are used as evidence against the owner. However, the comparisons are presently done by eye and this means that the evidence is not as robust as the Agency would like; since it is seen as subjective. The purpose of this study is to discover whether computer-based methods of pattern recognition and comparison could be used to provide evidence of a more objective nature, and, if suitable software does not currently exists, to recommend how best to develop it.

The project was conceived when the Agency became aware of work that had been published on the use of neural networks: a) to predict the production region of olive oils; and b) to detect adulteration of olive oils. It was felt that neural networks might be able to predict the source of an oil pollution event. Thus, this study initiated with the title "The application of neural networks to pattern recognition and comparison - a feasibility study", and the aim of making recommendations on how pattern recognition techniques, based on neural networks, could be used to identify the source of an oil pollution event. However, for reasons that are explained below, the scope of the study was broadened to include any computer-based method of pattern recognition or comparison, not just neural networks.

1.2 Objectives

The overall objective of the study, as stated in the original contract, was:

To determine the feasibility of using neural network techniques of pattern comparison for the purpose of oil identification.

The specific objectives were:

- To determine the feasibility of using neural network techniques of pattern comparison to compare the gas chromatograms (GC) or infrared spectra (IR) of spilled oils with those of standard oils and oils from possible sources of pollution.
- To investigate whether the use of neural network techniques of pattern comparison might increase the reliability and defensibility of oil identification relative to visual methods.
- To produce a Technical Report on the findings and, based on the findings, produce recommendations for further R&D.

At an early stage in the project it became apparent that neural networks may not provide the best means of achieving the project's overall objective. In consultation with the Project Manager, David Britnell, it was decided that the scope of the study should be broadened to include computer-based pattern comparison techniques other than neural networks.

1.3 Overview of the Study

The study consisted of several distinct exercises, as outlined below.

- Problem Appraisal - a detailed appraisal of the nature and technical complexities of the oil pollution problem was made with the aid of a specific example.
- Literature Search - included searches of scientific databases (EiCompendex, ISI, RSC and ChemoBro), plus general web searches.
- Literature Review - copies of key papers on both the pattern recognition/comparison and chemical analysis aspects of the study were acquired and reviewed.
- Contacts with Key Researchers - discussions were held with key researchers via telephone and e-mail.
- Software Review - details of available software were acquired from the suppliers and reviewed to assess their suitability for the task at hand.
- Development of a Potential Solution - an preliminary investigation was carried out, based upon a specific example, to test the viability of a potential solution to the source identification problem.
- Formulation of Recommendations and Conclusions.

All of these aspects of the study a described in detail in the following Sections.

2. NATURE OF THE PROBLEM

In order to gain a clear view of the nature of the problem, the authors requested details of a typical case and subsequently visited the Testing Laboratory in Fobney. Detailed discussions with the Contract Manager, David Britnell, and Senior Scientist, David Gazzard, brought to light some characteristics of the problem which cast doubt on the suitability of neural networks as a means to a solution. Thus, it was agreed that the scope of the project be revised to cover all computer-based methods of pattern recognition and comparison.

When comparing the problem of oil spills in rivers with other pattern recognition problems, like oil spills at sea or the identification of the production region of olive oil (which had been a source of inspiration for the development of this project), it was apparent that there were important differences. In the case of oil spills at sea and the classification of olive oils there is a wealth of data on the characteristics of the oils produced in different regions. Thus data exist on which to train neural networks to identify the characteristic fingerprints of oils from the different sources. Having trained a network it can be used to identify the source of an individual sample. In the case of oil spills in rivers the situation is quite different. Firstly, there is not a large database of samples from the potential sources, nor is it viable to compile such a database. Secondly, the potential sources differ greatly from one pollution event to another, depending upon the size and location of the river catchment in question. Thirdly, the fingerprints of the oils stored at the potential sources may change at intervals as stocks are replenished. Thus, the large amount of data necessary to train a neural network that is capable of reliably identifying a source simply does not exist, nor is ever likely to exist. It was therefore necessary to examine the basic characteristics of the problem from first principles and to determine what pattern recognition/comparison methods were most appropriate to its solution.

The essential characteristics of the problem are:

- the pollutant may be one of several types of refined oil (e.g. diesel, lubricating oil, etc.) each of which has its own characteristic GC/IR fingerprint;
- samples taken in the field may have weathered or degraded to the extent that their GC/IR fingerprints are very different from those of the original oil;
- weathering/degradation makes the matching of field and source samples more difficult;
- samples of potential sources are only acquired after the pollution incident has taken place;
- the type of refined oil responsible for a spill can be identified from key features in its GC/IR fingerprint;
- the identification of features that are unique to the source oil is more difficult, although some features are known to exist that are useful in this respect (e.g. the isoprenoid alkanes - pristane and phytane).
- very little use is currently made of information contained within the mass of poorly-resolved GC peaks or the unresolved complex mixture (UCM), although both appear to contain useful source-specific information that could be used for source identification;
- GC/IR fingerprinting methods are machine sensitive, making it important to analyse field and source samples on the same testing machine;

Our initial assessment of the problem was that its solution appeared to require four distinct tasks.

1. Collect and analyse field samples.
2. Compare the fingerprints of the weathered field samples with those of standard refined oils to identify the type of refined oil from which the field samples originated.
3. Identify potential sources, then collect and analyse samples.
4. Compare the fingerprints of the weathered field samples with those of the potential source oils to identify the actual source.

The fundamental issue which has to be resolved is how best to perform the comparisons required by items (2) and (4), bearing in mind that the fingerprints of the field samples have been modified by weathering and/or biodegradation. In fact, the problems posed by (2) and (4) are themselves quite different.

In the first case (2), the comparisons are made between the fingerprints of field samples and standard refined oils. Thus, it is similar to the problem of a crude oil spill at sea, where a large database of fingerprints of crude oils from different sources is used to identify the source. In addition, the fingerprints of the field samples also differ noticeably from those of their source oils due to the effects of weathering and biodegradation. Traditionally, computer-based methods of matching the fingerprints of field samples to those of standard oil types have been based upon statistical techniques, but more recently techniques based on neural networks have been introduced. The relative merits of these techniques are discussed later.

In the second case (4), the comparisons are made between the fingerprints of the field samples and those of potential source oils of the same type. The number of samples available from the field and the potential sources is very limited, so the problem reduces to one of comparing the distinctive features of the field and source fingerprints in the hope of achieving a clear match between the field samples and just one of the potential sources. This is the most important and most challenging aspect of the source identification problem. It requires the identification of sufficient source-specific features to enable a reliable match to be established, despite the confounding effects of weathering and biodegradation.

3. RESULTS OF SEARCHES

Searches were carried out using various scientific / engineering databases and the internet. The resulting lists of papers, research institutions, environmental consultants, software developers etc. were examined to determine which were directly relevant to the project. Copies of the most relevant scientific papers and details of commercially available software were acquired and reviewed. Key individuals were identified from the data collected and were later contacted by e-mail or telephone. These discussions resulted in further contacts and the accumulation of a world-wide view of the current state-of-the-art. The following sections describe the outcomes of the search process. Our interpretation of the materials acquired are given in section 4 'Findings' and are summarised in section 4.4 'Summary of Key Findings'.

3.1 Literature Searches

Literature searches were carried out using several databases, including the Institute for Scientific Information (ISI), Engineering Information (EiCompendex) and the Royal Society of Chemistry (RSC), in addition to more general internet searches and requests for information from various contacts established in the course of the study. These searches identified many relevant scientific papers and commercial publications, a full list of which is given in Appendix A, together with abstracts or summaries where available. Copies of the most relevant papers were obtained either from the collection of journals and conference proceedings held at Staffordshire University or from the British Library collection. These papers were reviewed and a list compiled of the most important ones, details of which are given in section 9 'Key References'. A list of literature on the analysis, composition, identification and solubility of petroleum fuels and oils is maintained by the American Petroleum Institute (API) on their web site (http://www.api.org/ehs/fuels.htm). A copy of the current list is given in Appendix A.

One fact that emerged from the literature search was that many papers were published on topics relating to oil characterisation and source identification in the period 1989-93, but that far fewer papers have been published since then. The activity during 1989-93 appears to have been generated by the Exxon Valdez oil spill in Prince William Sound, Alaska, in March 1989. No research appears to have been carried out into the problem of spills of refined products on inland waters, although there are a number of papers on the contamination of groundwater by oil products.

The key references which we have identified are of two types, those relating to fingerprinting techniques, especially gas chromatography, and those relating to computer-based techniques for pattern recognition or comparison. Some of the papers incorporate both of these aspects of the study.

3.2 Contacts

The following list gives details of the people with whom we had either e-mail or telephone discussions.

Andy Revill, CSIRO, Australia.
E-mail: Andy.Revill@marine.csiro.au

Harry van Enckevort, ESR, Wellington, New Zealand.
E-mail: Hvanenck@esr.cri.nz

Steve Grigson, Dept of Civil and Offshore Eng., Heriot-Watt University. (EUROCRUDE)
E-mail: S.J.W.Grigson@hw.ac.uk

Gerhard Dahlmann, BSH, Hamburg. (EUROCRUDE)
E-mail: gerhard.dahlmann@m3.hamburg.bsh.d400.de

David Holden, Hyprotech/AEA Technology, Didcot. (MatchFinder).
Tel: 01235 435544

John Wigger, Environmental Liability Management, Tulsa.
E-mail: jwigger@elmengineering.com

Meeji Ko, Infometrix, Woodinville WA. (Pirouette)
E-mail: info@infometrix.com

Steve Rowland, Department of Environmental Science., University of Plymouth.
E-mail: S.Rowland@plymouth.ac.uk

Sy Ross, S. L. & Steve Potter, Ross Environmental Research Ltd, Canada. (Oil Spill Exchange) E-mail: sy@slross.com & steve@slross.com

Ali Onder, Shell, Aberdeen.
Tel: 01224 882299

Andrew Tyler & Mathew Rymell, British Maritime Technology, Southampton.
Tel: 01703 232222

Ed Butler, BP, Sunbury.
Tel: 01932 763958

Bob Large & Peter Tibbitts, M-SCAN (Analysts), UK.
Tel: 01344 627612, E-mail: services.ltd@m-scan.com

Ian Kaplan, Global GeoChemistry Corporation, California.
E-mail: globalg1@idt.net

Allen Uhler, Battelle, Duxbury, MA (Environmental forensics).
E-mail:uhler@battelle.org

Andy Duller, School of Electronic Eng. and Computer Systems, University of Wales, Bangor.
E-mail: andy@sees.bangor.ac.uk

4. FINDINGS

4.1 Pattern Recognition Software

Several statistically-based pattern recognition packages have been developed that are capable of the identifying oil products from a library of standard types. Infometrix Inc. (http://www.infometrix.com) has developed a commercial software package called Pirouette that has been used on a range of environmental problems in the USA. A European consortium of researchers has developed a system, called EUROCRUDE, that uses a large database of crude oil fingerprints to identify the source of spills (Grigson & Baron, 1993, 1995; Sinclair and Grigson, 1996). Only one commercial package was found that used neural networks in its pattern matching procedures. This was MatchFinder developed by AEA Technology plc, which, according to the material on their web site (http://www.aeat.co.uk /pes/software/match.html), appeared to satisfy the requirements of the problem at hand, but unfortunately this product has been withdrawn from the market.

4.1.1. Pirouette by InfoMetrix

Infometrix, based near Seattle, is a leading supplier of chemometrics software. They have produced a number of technical articles, details of which are given in section 9.1 'Commercial Literature'. The company is actively involved in research and software development in the field of chemometrics. It was recommended to us by a leading researcher in the field, Dr. Allen Uhler, an environmental forensics specialist from Battelle Inc., Duxbury, MA. Their main product is Pirouette, a comprehensive chemometrics package available for Windows. A wide variety of prediction, classification, data exploration and pattern recognition methods are implemented, and the software is designed to be as flexible as possible rather than specialising in gas chromatographic pattern matching. It includes: a) hierarchical cluster analysis (HCA) and principal component analysis (PCA) for exploratory data analysis; b) k-nearest neighbours (KNN) and soft independent modelling of class analogy (SIMCA) for classification analysis; and c) partial least squares (PLS), principal component regression (PCR) and multiple linear regression (MLR) for regression analysis. A fuller description of Pirouette and its algorithms is given on the Infometrix website (http://www.infometrix.com/), together with a free demonstration version of the software. Due to the general multipurpose nature of Pirouette, the program has a fairly complicated interface with many functions, and is likely to take some time to master. It is not therefore the most appropriate tool for the task at hand.

4.1.2 EUROCRUDE

This system was produced by a consortium of researchers from Belgium, Denmark, Germany, Norway, Portugal and Scotland. The project was 50% financed by the Commission of the European Communities through its LIFE programme. EUROCRUDE was developed specifically for use in situations where only samples of the pollutants are available. That is, where it has not been possible to take samples from potential sources, such as oil tankers. Consequently, the system attempts to identify the oil field from which the crude oil originated. It is therefore based upon a large database of fingerprints of crude oils that are typical of the different oil fields. The pattern recognition system is based on the use of 15 key biomarkers divided into three groups (i.e. five triterpanes, five steranes and five aromatics). The peaks within each group are normalised and the degree of match between a sample and a potential

source is determined separately for each group. The results from the three groups are then used to identify the specific source. Various statistical methods of matching the peaks were tested, including absolute ranking, T-test, k-nearest neighbours (KNN), soft independent modelling of class analogy (SIMCA), a procedure based upon the *F*-statistic and the Euclidean distance. According to the most recent paper (Sinclair and Grigson, 1996), the method finally adopted was the Euclidean distance after normalisation.

EUROCRUDE was first put to the test on a real case when Environmental Science and Research Ltd. (a Crown Research Institute) in New Zealand used it to investigate an oil slick in Lyttleton Harbour. They reported (See http://www.esr.cri.nz/services/analytical/success-stories.html) that "it passed with flying colours", and concluded that their successful prosecution was the result of a "technically qualified team with good forensic methods and lots of good experience under cross examination."

4.1.3 MatchFinder by AEA Technology

MatchFinder is a chromatographic profile matching program produced by AEA Technology at Harwell. It was first developed by AEA jointly with Perkin-Elmer as part of the European Union IT programme ESPRIT (Phase 2) Application of Neural Networks for Industry in Europe (ANNIE), which ran from 15/11/88 to 14/11/91. A description of the algorithm used and results of tests are given in Mason *et al.* (1992). The system was further developed for commercial use after completion of the ANNIE programme. A major fuel controversy at the 1995 Brazilian Grand Prix drew attention to MatchFinder, with interest expressed by McLaren, Williams and Benetton racing teams. However, our contact at AEA Technology informed us that MatchFinder is not presently being marketed and the company has no plans to develop it further, although it is still features on their website at http://www.aeat.co.uk/pes/software/match.html. Nevertheless, MatchFinder was an interesting package in that it used neural networks to match corresponding peaks with respect to their retention times (Mason *et al.*, 1992). When comparing intensities, it used correlation analysis to determine the degree of fit between the matching peaks. Koussiafes and Bertsch (1993) tested the ability of MatchFinder to match the GC profiles of simulated arson samples and a library of standard accelerants. They concluded that the strength of the algorithm was its adjustment of shifts in retention time and that further investigation into the technique was warranted. Perhaps the outcome of this study was the reason for MatchFinder being withdrawn from the market.

4.1.4 Other commercial software packages

There are several other general pattern matching packages similar to Pirouette, although discussions with our contacts seem to indicate that Pirouette is the best. They include Sirius from Pattern Recognition Associates (http://www.main.com/~pra/sirus.htm) and Unscrambler from CAMO (http://camo.no/Products/TheUnscrambler7.html), which, like Pirouette, use the PCA, PCR, PLS and SIMCA techniques. There are also some very powerful general purpose statistical analysis packages, like Statistica by StatSoft (http://www.statsoft.com), that offer a very comprehensive range of analytical and visualisation techniques that could be used for the matching of oil fingerprints. However, such packages are so comprehensive that the time taken to become fully familiar with them would be unacceptable in this case.

4.2. Recent Pattern Recognition Research.

Several researchers have developed experimental software for use in environmental forensics, or have used general purpose statistical and/or neural network packages to investigate their potential uses in environmental forensics. A feasibility study by Wigger and Torkelson (1997) assessed the use of two statistically based algorithms as aids to the interpretation of hydrocarbon fingerprint data, and these were later applied to a case study (Wigger *et al.*, 1998). Several papers have been published (Elling *et al.*, 1997; Long *et al.*, 1991; Welsh *et al.*, 1996; Duller *et al.*, 1996) based upon studies into the use of neural network for the analysis and interpretation of chemical fingerprints. In most of them the accuracy achieved by the networks is compared to that achieved by commonly used statistical techniques (e.g. KNN and SIMCA). The results of these studies are outlined in the following two sections.

4.2.1 Statistical approaches

Environmental Liability Management (ELM) Inc. is an environmental consulting and engineering company based in Tulsa, Oklahoma. John Wigger from ELM and Bruce Torkelson from Torkelson Geochemistry, Inc. have published two papers on the problem of source identification of spilled oil (Wigger & Torkelson, 1997; Wigger, Beckmann, *et al*, 1998), the second of which was produced in collaboration with staff at Amoco.

Wigger's method was developed using a database of about 60 reference oil samples of many different types. In their original study (Wigger & Torkelson 1997), a set of 71 compounds (i.e. chromatographic peaks) was chosen to characterise each sample, but this was later increased to 89. The set included n-alkanes, olefins, iso-alkanes, naphthenes, isoprenoids, aromatics, polynuclear aromatics and oxygenates, but the basis on which individual compounds were chosen is not clear. However, the method is not dependent on the set chosen, so it would be possible to use the same method with a different set of characteristic compounds. Furthermore, it would be possible to choose a 'focused data set' containing only those peaks within a given range of interest.

Two algorithms were developed, one to determine the degree of similarity among different hydrocarbon samples, and the other to model the evaporation portion of the weathering process in gasoline. In the first, the cross-correlation coefficient between the two data sets (i.e. the sample and potential source data after normalisation of the peaks) was used as the measure of relationship between them. Note that although a large database was used for the development of the method, it does not rely on the database and can be used to compare any two oils. The second algorithm, which predicts the effects of evaporation, was developed from the results of controlled evaporation experiments. Using this algorithm, source oils can be artificially 'evaporated' and then compared to weathered field samples or, conversely, the effects of evaporation on a field sample can be artificially removed for comparison with a potential source. It was suggested that further controlled experiments would enable the technique to cope with any particular form of weathering and/or biodegradation.

Wigger *et al.* (1998) concluded that the method elucidated subtle differences and similarities in the chromatograms and provided a repeatable, user independent, quantifiable measure. However, in the discussion of results it was stated that although the method proved extremely helpful in quantifying similarities and differences among samples, there were several

indications that the technique must be used in conjunction with experienced visual interpretation.

Lavine *et al.* (1998) use discriminant analysis to classify the GCs of weathered and unweathered jet fuels into one of six types (i.e. JP-4, Jet-A, JP-7, JPTS, JP-5 and AVGAS). The study was based on a training set of 271 unweathered samples and a prediction set of 31 weathered samples, and was carried out in four distinct steps: (1) peak matching, (2) outlier analysis, (3) feature selection, and (4) classification. The most interesting feature of this study was that it used a genetic algorithm to optimise the feature vector used for the classification. That is, the task of determining the best subset of peaks (20 in this case) on which to base classifications was achieved through the use of a genetic algorithm. The published results indicated 98.8% successful classification of the training set (i.e. unweathered samples) and 100% successful classification of the prediction set (i.e. weathered samples). However, it should be noted that the weathered samples were taken from groundwater sources, so the degree of weathering was relatively slight. Unfortunately, several aspects of this study cast doubt on the validity of these apparently outstanding results.

Firstly, the outlier analysis eliminated 15 of the 271 samples from the training set because principal component analysis indicated that they did not fit well within their fuel type. Thus they were eliminated not because they were incorrectly classified, but because they were extreme examples of their type. Although this might be quite sensible for the purpose of developing the classification model, it is not acceptable to define the model's overall performance on the remaining 256 'consistent' samples, which is what was done. Furthermore, the prediction set was dominated by the most easily classified fuel types, JP-4 and AVGAS (making up 24 of the 31 samples in the set), thus casting doubt on the validity of the result of the performance test.

Secondly, the fitness function used for the genetic algorithm was based on prior knowledge of the fuel type of the samples, thus the training process was effectively based on supervised learning. Thus, bearing in mind the relatively small size of the training set (256 samples), it is likely that the optimised feature vector was overfitted to the data. This problem could have been overcome by the use of cross validation, but this was not done. In addition, some rather subjective procedures were used to filter the data prior to the feature selection stage. In our view, these were unnecessary and possibly detrimental to the development of a sound model.

Despite these weaknesses, the study is worthy of mention since it demonstrated the potential of genetic algorithms to optimise the feature vector. However, it is worth noting that there are several other techniques that can be used for this purpose (e.g. techniques based on neural networks or information theory).

4.2.2 Neural networks approaches

Elling *et. al* (1997) describe the development of a hybrid artificial intelligence system that automates the process of validating routine GC data and diagnosing instrument malfunctions. It combines a rule-based expert systems approach with a neural networks approach to produce a hybrid system, known as an expert network, in which a set of rules elicited from experts is represented in the form of a neural network. Once the rule-base has been translated into an expert network format, its performance can be enhanced using relevant data to train the network, in the same way that one would train a supervised-learning neural network. It is

claimed that the key advantage of expert networks is that they preserve the knowledge representation and explanation capabilities of expert systems, whilst providing the learning capabilities of neural networks. However, we are not entirely convinced by this claim, because it fails to recognise the inherent weaknesses of the rule-based approach to reasoning under uncertainty. Nevertheless, the system developed by Elling *et. al* (1997) undoubtedly provides a very effective means of validating GCs prior to use in data processing systems. It could serve as a valuable front end to a pattern matching system by guarding against false conclusions being drawn from erroneous data. Basically, the system performs a two step process that attempts to emulate the behaviour of experts. First, it analyses the overall appearance of the GC and identifies any symptoms of abnormality, such as: an elevated, oscillating or noisy baseline; a lack of peaks; or abnormally shaped peaks. It then uses all of the identified symptoms to diagnose the underlying fault in the sample or instrument.

A study by Long *et al.* (1991) examined the potential of neural networks as pattern recognition tools for use on chromatographic data. They compared the ability of neural networks (NN) to identify the GCs of seven different jet fuels with those of two commonly used statistical methods, namely k-nearest neighbour (KNN) and soft independent modelling of class analogy (SIMCA). The neural networks used were based upon the standard back-propagation (i.e. supervised-learning) algorithm and had a single hidden layer. The networks' parameters and architecture were optimised to maximise correct classification rates. Tests were carried out on two different data sets: one consisting of 48 features of 126 chromatograms of water-soluble fuel fractions; and the other consisting of 131 features of 154 chromatograms of neat oils. Both sets of tests were cross-validated by dividing the data into roughly equal training and testing data sets. The results showed that the percentage of correct classifications achieved by KNN, SIMCA and NN were 79%, 70% and 98% respectively on the water-soluble fraction test set, and 82%, 94% and 89% respectively on the neat oil test set. Clearly, the neural networks outperformed the two statistical methods in one test but not in the other. However, two points are worth noting about these tests: 1) most of the misclassifications involved three fuels (Jet-A, JP-5 and JP-8) which are so similar that on occasions a fuel has been shown to satisfy the specification of two of them; and 2) the neural networks were over parameterised given the small size of the training data sets (i.e. over 1000 parameters compared to less than 100 training cases). Although the use of cross validation would have overcome much of the danger of overfitting the training data, the high degree of over parameterisation could have resulted in some overfitting of the test data. This being the case, the true performance of the neural networks may have been overestimated. The paper concluded that neural networks had been shown to be a useful pattern recognition tool for the classification of chromatographic data from jet fuels, which, even given our comments, was a reasonable conclusion.

A study by Welsh *et al.* (1996) evaluated several computer-based classifiers as potential tools for pharmaceutical fingerprinting using normalised data derived from HPLC trace organic impurity patterns. Artificial neural networks (ANN) were compared to two standard chemometric methods, k-nearest neighbour (KNN) and soft independent modelling of class analogy (SIMCA), as well as a panel of human experts. The number of inputs to all three models was varied to achieve optimum performance, and the number of nodes in the hidden layer of the neural networks was also optimised. Three sets of input data were tested, two containing 22 and 46 key peaks and the third containing all 899 data entries extracted from the chromatograms. The best performance achieved by each method, expressed in terms of the

percentage correct classifications, was ANN-46 (93%), SIMCA-46 (87%), KNN-46 (85%) and 'human experts' (83%). However, there were aspects of the project which cast doubt on the reliability of the results. Firstly, the number of samples used for the tests was 253, whereas the topology used for ANN-46 meant that it had 2066 parameters. Thus, despite the use of cross-validation in the training phase, it is likely that the neural network models were overfitted to the data. Secondly, the panel of 'human experts' was made up of graduate and post doctoral students from the Department of Chemistry at the University of Missouri. Since no mention is made of the extent of their experience at analysing GCs for the specific purpose of this study, the validity of the label 'human experts' is questionable. Nevertheless, the study does indicate that neural networks may be capable of outperforming standard statistical techniques on this type of pattern recognition task.

Duller *et al.* (1996) examined the ability of two unsupervised-learning neural networks to classify crude oil fluorescence spectra. The two networks tested were Self-Organising Feature Maps (SOFM) and Adaptive Resonance Theory (ART-2a). Their accuracy was assessed relative to each other and relative to results achieved by a human interpreter. Differences were apparent in cluster memberships and configurations derived by SOFM and ART-2a. It was concluded that the geochemical significance of these differences was unclear, but that SOFM results may be more meaningful and of greater practical significance to geochemists. A comparison of the benefits of SOFM and ART-2a listed several benefits of ART-2a that relate to its training process. However, we are of the opinion that the relatively short times involved in the development of neural network or statistical classifiers of any kind is of little overall significance, especially when compared to the time taken to construct and validate the necessary database. The most important factor in our view is the accuracy and reliability of the final model. The overall conclusion of this study was that the neural networks tested have the potential to reduce the tedium but not to de-skill the task of analysing the results obtained from geochemical surveys.

Finally, a study carried out by Walley *et al.* (1998) for the Environment Agency demonstrated the ability of Self-Organising Maps (SOM) to identify patterns in biological and environmental data for the purpose of river quality classification. This study has since been extended to a further study (National R&D Project E1-056) which aims to improve the pattern recognition capabilities of the SOM and to use it to identify the pollutants, or types of pollutant, that are the cause of river quality degradation. The results of this project may produce pattern recognition software that could be of value in the identification of the type of oil responsible for a pollution event.

4.3 Analytical Methods

4.3.1 Comments on current protocols

The protocol currently used by the Environment Agency for producing data on which to fingerprint oil products consists of solvent extraction, followed in some cases by drying and then GC with FID detection. This produces a trace consisting mainly of n-alkanes with some isoprenoid alkanes, and a small component of aromatic hydrocarbons in some samples. By using internal standards in conjunction with an overall hydrocarbon standard, a process of visual comparison can produce a match which may then be used as evidence to support a legal case against the suspected polluter. This is done by first matching the fingerprint of the sample with a reference sample, to identify the type of pollution, and then matching it to

samples from potential polluters, to identify the specific source. This approach is by no means uncommon, but it has some serious shortcomings.

The first of these is comparability of process. A comparison of the Methods Manual (NRA Thames Region Laboratory, March 1993, Issue 2) and the report on the case study provided by Environment Agency (M-Scan Report 9804/10265, 10 Sept. 1998), reveals that quite different extraction solvents (pentane or DCM) may be used, along with different extraction systems (ultrasonics or shaking). Furthermore, the manual does not stipulate the type of column to be used. This lack of comparability between the protocols of different laboratories, whilst recognising that both are equally valid, makes comparisons between traces produced by different laboratories very unreliable. The implication of this for the proposed pattern recognition system is that such a system will only operate effectively if used on traces produced to the same protocol. This point will be of particular importance if the proposed pattern recognition system is developed using a centralised library of reference patterns. An interesting spin-off from the neural network study by Welsh *et al.* (1996) was that it exposed evidence of significant variations in the HPLC chromatograms across LT manufacturers, across three HPLC columns and, for one manufacturer, across different lots. The paper states that the extent of column-to-column variations was particularly noteworthy in that all three columns had identical specifications with respect to their stationary-phase characteristics and two of the three columns were from the same vendor. We found no evidence of similar variations in GC in any of the papers we reviewed. Thus, it appears that the use of HPLC would add another dimension to the problem of analytical variability (i.e. significant variability within protocols in addition to that between them).

The second shortcoming is that of sensitivity and subjectivity. Considerable difficulty can be encountered in discriminating between certain kinds of sample. For example, JP8, 28 second fuel oil, kerosene and AVTUR all give very similar profiles, largely differentiated by some instances of relative peak height variation. The very similar nature of the samples makes visual discrimination between them very subjective and unreliable, especially if there are differences in extraction and analysis protocols. Computer-based methods of pattern matching also have difficulty discriminating between these types of fuel oil. The neural networks study by Long *et al.* (1991) found that all systems tested, both neural and statistical, had difficulty in discriminating between jet fuels Jet-A, JP-5 and JP-8. For more complex extracts, like creosote or unrefined oil, the GC traces produce much more complex patterns, which show considerable variation between oil and creosote types. In addition, their aromatic content is likely to be higher, thus adding more confusion, but also extra information which may be useful for discrimination.

A third problem which will arise for environmental samples, particularly ones representing longer term pollution, is that the sample may contain multiple pollutants of a similar type. (e.g. two or more mixed hydrocarbon products, possibly of differing age and therefore state of weathering). This problem could not readily be addressed through the fingerprint comparison system in current use.

4.3.2 Findings from the literature review

Our literature review has shown that the need to address these problems is being recognised world-wide. Two approaches are being taken.

Firstly, protocols are being reviewed and refined, making them more consistent within individual studies, in conjunction with quantitative analysis of the chromatograms so produced. An example of this was presented by Wigger *et al.* (1997, 1998) who focused on producing a reference set of profiles from a tightly defined GC FID protocol for 60 known hydrocarbon samples and the identification of 89 key features within them. A similar approach was taken by Grigson and Baron (1993, 1995) to produce a reference set for 300 crude oils based upon 56 key features. The analytical technique used in this case was GC-MS.

Secondly, there is clearly an emerging hunt for an improved range of either chemical analyses or data treatment regimes, or both, to improve the diagnostic power of environmental forensics.

A comprehensive review of chemical fingerprinting techniques by Wang and Fingas (1997) concentrated on methods used for the characterisation of petroleum hydrocarbons, the identification of oil spills and the assessment of environmental impacts. It recognised the value of analytical techniques such as GC-FID, GC-MS, HPLC, IR, UV and FL spectroscopy, for a variety of purposes, either to determine total petroleum hydrocarbons (not applicable here) or to determine individual components or a set specific to petroleum hydrocarbons (fingerprint). Although it was written from the perspective of experience in the United States and Canada, mainly based on ASTM and EPA methods, it nonetheless provides a comprehensive and up-to-date review of the field. Although it focuses on new trends and developments in oil analysis methods, it also covers the effects of weathering and biodegradation, and the use of biomarkers for source identification. Neural networks are not mentioned, but it is concluded that statistical methods of pattern recognition, based on principal component analysis and discriminant analysis, have demonstrated the ability to identify the source of an oil. It is suggested that further developments in this area will lead to even greater improvements in source identification. The paper includes a substantial bibliography (132 references), and can thus be regarded as the primary reference point for the current 'state-of-the-art' and background information in this field.

A study of the relative performance of ASTM methods in the laboratory and field following the Exxon Valdez disaster (Hendrick and Jadamec, 1991) found that capillary gas chromatography appeared to be more reliable than either HPLC, TLC, IR or FL. It concluded that a multi-method approach is the most efficient and reliable way to positively match a spilled (crude) oil to a common source.

An investigation into the use of chemical oxidation of unresolved complex mixtures (UCM) to yield gas chromatographically resolvable compounds (Revill *et al.*, 1992) demonstrated how additional information for the identification of source crude oils can be derived from UCM analyses.

A good general overview of the interdisciplinary approach to the unravelling of environmental liability at contaminated sites in the USA is given by Stout *et al.* (1998), and material produced by Arthur D Little, Inc. (http://www.arthurdlittle.com) outlines a set of advanced chemical fingerprinting methods which they describe as a critical set of tools for companies facing liability claims.

4.3.3 Issues raised

The main issue raised by this aspect of the study is: do we need to seek improvements in chemical analysis techniques in order to discriminate fingerprints through a pattern matching system?

Given that any cross sample comparison will require absolute consistency of chemical extraction, treatment, and analysis, in order to provide dependable data, will the pattern matching system be capable of producing reliable results using existing chemical analysis systems, or will it require a more refined system?.

Our view is that the pentane extracted, GC-FID analysed system, with no intermediate clean up or derivatisation processes, probably offers enough data in the minor peaks on which to work. We see no merit in using additional intermediate stages since these could: a) reduce the data, and hence the useful information, by cleaning it out; b) introduce another area for potential variation in protocols; and c) increase the time and cost of sample processing.

We also feel that it would be highly desirable for samples to contain internal standards, and to be run against standard hydrocarbon mixtures, such that the chemical position of the parts on the chromatogram to be matched can be confidently and consistently identified.

Any pattern matching system based on relative peak height will need to be presented in a way that demonstrates an understanding of the behaviour of individual components under various extraction or weathering conditions. To do this the patterns will need to be presented in the context of compounds in the profile which make up the pattern being discussed.

It may be that for complex, parent oil or creosote type samples, an additional analysis, which brings out the aromatic fingerprints, would add to the discrimination process. A DCM or Pentane extract subjected to HPLC fluorescence detection could be considered for this function, but measures would have to be taken to minimise the analytical variability mentioned earlier (4.3.1). In other cases where it is difficult to find sufficient distinguishing features to form a confident match, an analysis of the UCM based upon the method described by Revill *et al.* (1992) could be used to draw out additional features.

Finally, any pattern recognition system must be robust in the matches it forms and sensitive in its ability to differentiate between similar oils. We believe that follow up work to test these two aspects of the systems on a wide range of oil types of varying ages will be essential to prove the technique. This could usefully include an investigation into variations between different laboratories using the same protocol, and between items of equipment of the same specification.

4.4 Summary of Key Findings

- Two software packages for the matching of samples to source oils are currently available. They are: EUROCRUDE which is designed specifically for the identification of crude oils from a library of source characteristics; and Pirouette which is designed for more general use. Both are based on statistical methods of pattern comparison. A third system MatchFinder has been withdrawn from the market.

- Other GC pattern matching software has been developed and tested by researchers, but it is not commercially available.
- Research and development work to date has been primary based on the identification of crude oils, although some work has been based on the identification of contaminants in groundwater.
- Neural networks have not yet been used for pattern comparison in commercially available software, but they were used by MatchFinder to correct retention time shifts prior to matching GC peaks.
- Neural networks have been used by several researchers for the recognition of GC, HPLC and IR fingerprints. Some have compared the performance of the neural networks with those of standard statistical methods and domain experts, and found that the neural networks outperformed both. However, the validity of these results is questionable.
- Gas chromatography appears to be the most appropriate and most reliable analytical technique, although there may be a need for tighter protocols to ensure that like-for-like comparisons are made.
- For the purpose of pattern comparison the GC profiles are normally reduced to a set of peaks of key compounds and then normalised to 'eliminate' the effects of weathering / biodegradation. The latter appears to have worked satisfactorily, but more sophisticated methods of accounting for weathering have been developed.
- No research or development work has been done specifically on the use of pattern recognition for source identification of oil spills on inland waters.

5. TOWARDS A SOLUTION

5.1 Overview

As stated earlier, the identification of the source of an oil spill on an inland water requires a two-stage approach: 1) identification of the type of oil product responsible for the spill; and 2) identification of the specific source from which it originated.

The problem posed by the first stage (i.e. type identification) is similar to that presented by a crude oil spill, and can be tackled in the same way to that used for EUROCRUDE. It requires a database of typical fingerprints of the whole range of oil products based on a predefined set of indicator compounds. Statistical or neural network pattern recognition techniques, with adjustments for the effects of weathering, could then be used to identify the oil type responsible for the spill. It may be possible to use existing commercial software (e.g. Pirouette) to do this, but it would be better to develop software specifically for the task. In either case, a database of typical fingerprints would have to be compiled. Thus the first stage of our proposed solution does not require the development of new techniques, but it does require the testing of statistical and neural network techniques to determine which performs best on the particular pattern matching task. This will require the compilation of a comprehensive set of test cases.

The second stage of the proposed solution is more challenging, since it requires the development of new methods. The amount of information provided by the main (or highly resolved) peaks of a GC is, in our view, insufficient for the reliable identification of the source of a spill. Our initial appraisal of the problem indicated that there are distinctive features within the bands of minor peaks lying between the main peaks that could provide the necessary additional information. To test the viability of this idea we carried out an exploratory investigation based on set of typical data provided by the Environment Agency.

5.2 Exploratory Investigation

5.2.1 Introduction

This investigation was based on the principle that if human beings can identify the source of an oil spill from the patterns of minor peaks within a GC trace, then it is possible for computers to do so. Conversely, if humans cannot do this, then it is highly unlikely that a computer will be able to do so. Thus, the idea was tested out on 20 human volunteers using a visual pattern matching exercise.

5.2.2 The data

The data for the study were supplied by the Environment Agency's Project Manager, David Britnell, in the form of digital GC outputs (i.e. peak intensities and peak areas against retention time). The GC outputs of nine samples were used in the study: three diesel-contaminated field samples (one water sample, one soil sample and one sample of oil from the water surface); one sample from the actual source of the spill; and five other samples of unweathered diesel. The digital files were used to produce graphical representations of the GC outputs.

Each graph had a similar profile in terms of its major peaks, although noticeable differences did exist between the field and source samples due to the effects of weathering. In addition, there may have been difference due to the fact that the GCs of the five non-source samples were produced by the Environment Agency's laboratory at Fobney, whereas the GCs of the other samples were produced in a consultant's laboratory.

5.2.3 Method

Sections of each graph were extracted by eliminating the major peaks and retaining the minor peaks between them, as illustrated in Figure 5.1. The choice of 'cut-off points' for each section was made by eye and corresponding sections were matched by their retention times, which, although similar, were not identical owing to experimental variations from test to test. The corresponding sections from the GCs of the nine samples were grouped together for the purposed of the visual pattern recognition exercise. Since it was sometimes difficult to define the precise end points of the sections, participants in the exercise we told to place less weight on the end points and focus on the main body of the patterns. In some cases, the vertical scales of the graphs differed considerably, due to the effects of weathering. Since it was the pattern in the GC trace that we wished to focus on, we removed all scales to produce a series of lines connecting points on an otherwise blank sheet (i.e. a collection of patterns rather than graphs was produced). If it could be shown that humans could successfully match sections of the GCs of the field samples with the corresponding sections of the true source oil, rather than any other 'potential source', then it would follow that these small sections do indeed contain valuable source-specific information. In addition to asking our 20 volunteers to match the patterns, we asked them to assign a 'degree of certainty' (i.e. on a 0 - 10 scale) to their chosen matches. Full details of the visual pattern matching exercise and the instructions given to the participants are given in Appendix B.

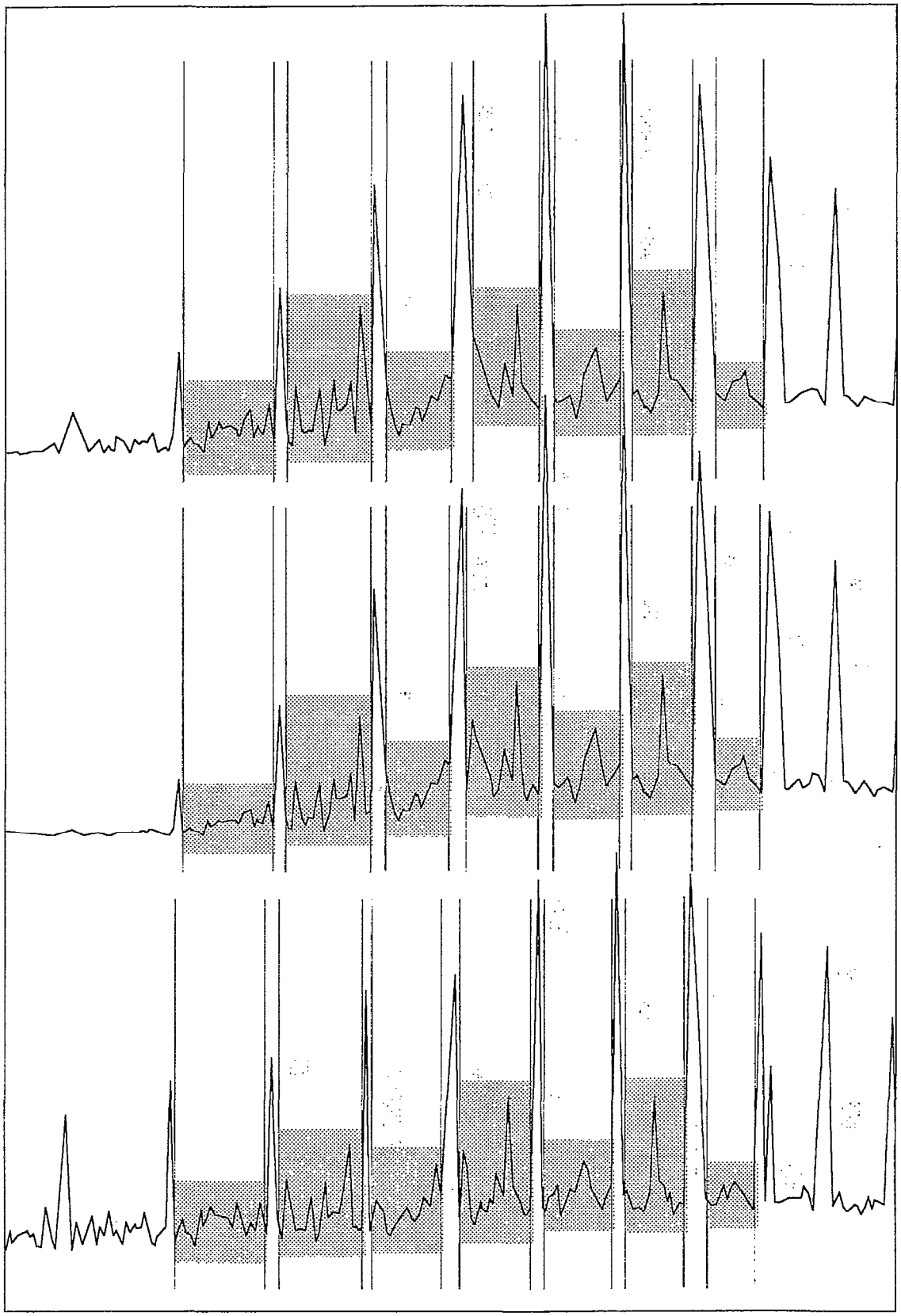


Figure 5.1 Gas chromatogram data plotted for the true source (top), water sample (middle) and a 'false' source (bottom), covering retention times from 10 to 25 minutes. The shaded areas indicate the sections between the major peaks used for the pattern matching exercise.

5.2.4 Results

A summary of the results of the visual pattern matching exercise is given in Table 5.1. In addition, Figure 5.2 shows the percentages of participants who correctly identified the true source of each of the three field samples from the patterns in the seven sections of their GCs. Figure 5.3 shows the same information plus the numbers of participants who incorrectly identified the source as one of the 'false sources'. Full details of the results are given in Appendix C.

Table 5.1 Summary of results of the visual pattern matching exercise

	Identification of Source Percentage Correct			Average Degree of Certainty of Match (0-10 scale)					
				Correct Responses			Incorrect Responses		
Group No.	Water Sample	Surface Sample	Soil Sample	Water Sample	Surface Sample	Soil Sample	Water Sample	Surface Sample	Soil Sample
1	100	85	55	8.00	6.47	5.73	-	3.67	4.89
2	95	90	25	7.21	7.78	4.20	7.00	6.50	4.40
3	100	100	90	9.15	7.75	7.06	-	-	6.00
4	20	0	5	4.75	-	4.00	4.88	5.25	4.84
5	95	95	75	8.53	7.95	7.27	5.00	5.00	3.60
6	100	95	95	9.55	9.32	8.47	-	7.00	7.00
7	100	100	100	9.00	7.80	8.05	-	-	-
Avg.	87.1	80.7	63.6	8.03	7.85	6.40	5.63	5.48	5.12
Avg. (excl.4)	98.3	94.2	73.3	8.57	7.85	6.80			

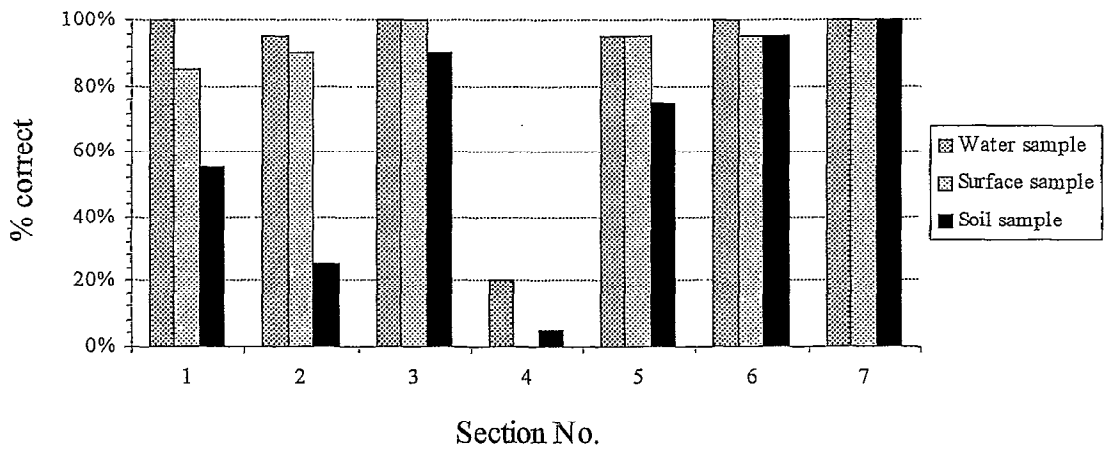


Figure 5.2 Histogram showing the percentage of the 20 participants who correctly identified the true source of the field samples from the patterns in each of the seven sections of the gas chromatograms.

The results clearly show that the participants were able to identify the true source with a fairly high degree of certainty in six out of the seven cases presented, although greater success was achieved with the water and surface samples than with the soil samples. Thus, the pattern characteristics of six of the seven GC sections were each sufficiently distinct to permit identification of the true source with a fair degree of confidence (i.e. with certainty levels normally greater than 7.0). By combining the results of all seven cases the true source could be identified with overwhelming certainty. In the one case where the true source was not identified (i.e. section 4), more than 50% of participants identified the source as 'false source 1', but with certainty levels normally less than 5.0, and several participants indicated that it was not possible in this case to identify a match with any of the potential sources. Clearly, the pattern characteristics of this section of the GC of the true source are not sufficiently different from those of the other potential sources to provide a reliable means of discriminating between them. If we concentrate only on sections where at least 90 percent of participants identified the true source, then:

- six of the seven sections of the water sample's GC were matched to the true source with an average certainty level of 8.57;
- five of the seven sections of the surface sample's GC were matched to the true source with an average certainty level of 8.12; and
- three of the seven sections of the soil sample's GC were matched to the true source with an average certainty level of 7.86.

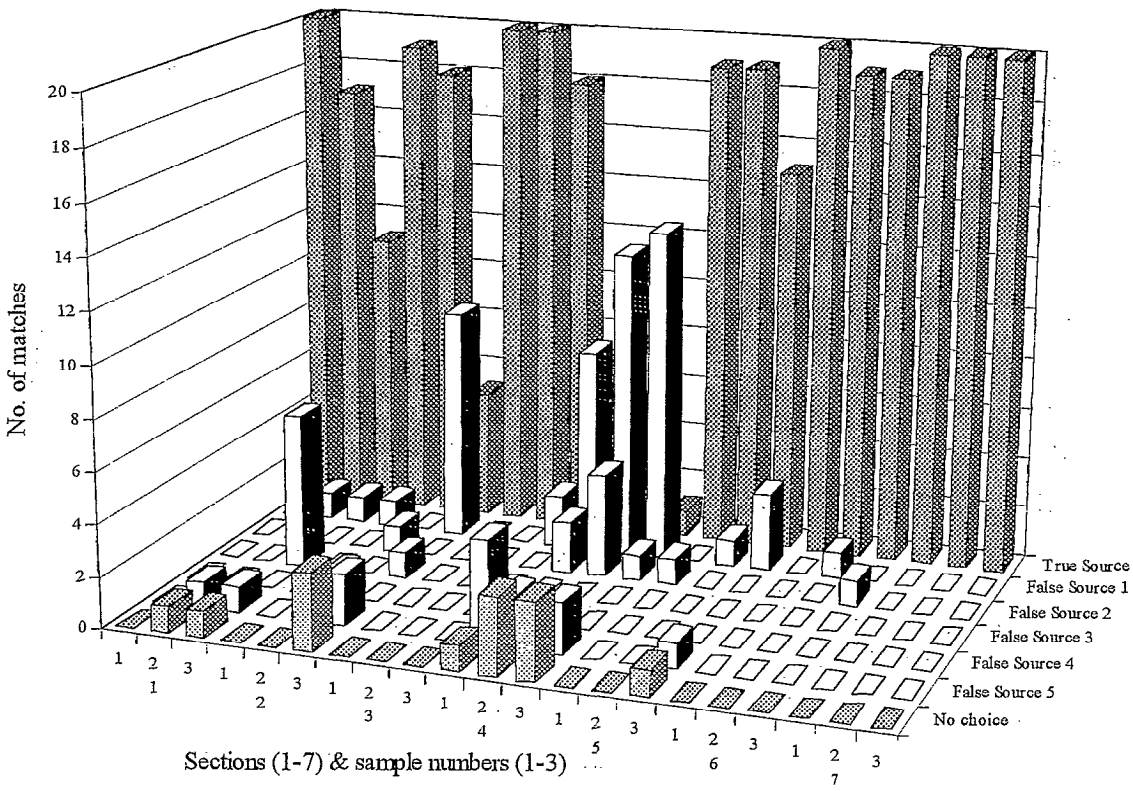


Figure 5.3 Histogram showing the distribution of matches made between samples and sources for each of the seven section of their GCs. Where participants were unable to make a match this was categorised as 'no choice'. The water, surface and soil samples are labelled 1, 2 and 3 respectively.

One section (section 7) of the GCs of the water, surface and soil samples was matched with the true source by all participants, with average certainties of 9.0, 7.8 and 6.4 respectively. Two other sections (3 and 6) of the water, surface and soil samples were matched with the true source by at least 90% of the participants. Thus, three sections (3, 6 and 7) had patterns that were clearly unique to the true source, even after the effects of weathering/ biodegradation.

5.2.5 A computer-based approach

Exploratory studies were carried out to determine how the visual pattern matching achieved by our volunteers might best be replicated by a computer-based system. We concluded that it could be achieved using a statistically based pattern matching approach in which the effects of retention time shift and weathering are first minimised by appropriate rescaling of the axes. Our study indicated that these could be achieved by: a) rescaling retention times uniformly between benchmarked major peaks; and b) rescaling intensities as a linear function of retention time. The peaks would then be matched in terms of their retention times, and the GCs divided into *n* sections of minor peaks between adjacent major peaks (e.g. the n-alkanes), as in the above study. The field samples would then be compared with the potential sources in terms of the degree of fit between their patterns in corresponding GC sections. The degree of fit between them could be defined in terms of their Euclidean distance or their correlation coefficient (as used by Wigger *et al.*, 1998). The system would have to be designed to cater for the situation where one sample has more peaks in a given section than the other, leaving one or more unmatched peaks. This commonly occurs when a peak is so weak that it is identified in one GC but not in the other. The system could either: a) not utilise sections in which there are unmatched peaks; or b) ignore unmatched peaks when doing the comparison, but record how many were found. The matches found in each of the sections of the GC would then be ranked in order of their degree of fit (i.e. Euclidean distance, correlation coefficient or other appropriate statistic) and used to produce an overall conclusion (i.e. based on all *n* sections), possibly with an associated degree of confidence.

5.2.6 Peaks heights versus peak areas

It should be noted that this preliminary study was carried out using peak heights only, and that it is possible that had peak areas been used, as has been done by several other workers, the results might have been even better. Clearly, both types of data contain useful (albeit related) information and each could contribute valuable evidence towards the identification of the true source. Thus, we recommend that any system developed should use both peak heights and peak areas to match the field samples to potential sources.

6. DISCUSSION AND RECOMMENDATIONS

The study determined the current state-of-the-art and highlighted a number of areas where it is appropriate to offer advice and make recommendations for further action. The following sections discuss all of these matters and make detailed recommendations where appropriate.

6.1 State-of-the-Art

Although a few computer-based pattern recognition systems for the identification of the source of oil spills are commercially available, none of them is directly applicable to the problem at hand. They all relate to crude oil spills at sea and the pollution of coastal areas, a problem which, although similar to the one at hand, presents a noticeable different data analysis problem, as explained in section 2 'Nature of the Problem'. Some research studies have been carried out which relate more closely to the problem of identifying the type of refined oil responsible for a spill on inland waters, but no evidence has been found of the development of systems for the identification of the source of the oil spill. Most systems have used statistical methods of pattern matching, although more recently systems have been developed using both supervised-learning neural network (e.g. backpropagation MLP) and unsupervised-learning neural networks (e.g. SOMF and ART). However, there is not yet sufficient evidence to show whether neural networks are any better than the standard statistical methods, or which of the various neural networks is best suited to the classification of oil fingerprints.

6.2 Analytical Methods and Protocols

According to an article on 'Advanced Chemical Fingerprinting' by Boehm *et al.* (http:// www.arthurdlittle.com) of Arthur D. Little, Inc., standard analytical protocols in the U.S. have proven inadequate for the advanced chemical fingerprinting necessary for pollution liability cases. Thus, there appears to be a need for improved analytical protocols if the full benefits of advanced chemical fingerprinting are to be realised. However, this is an issue that lies outside the scope of this report. The study by Hendrick and Jadamec (1991) mentioned earlier (4.3.2) concluded that GC appeared to be more reliable than either IR, FL, HPLC or TLC, and that a multi-method approach is the most efficient and reliable way to positively match a spilled (crude) oil to a common source. Welsh *et al.* (1996) encountered variability problems with HPLC (see section 4.3.1). These facts, together with our knowledge and experience of the various methods, lead us to the conclusion that GC is the most appropriate method to use, but that cases may arise where it does not provide sufficient evidence on which to draw a firm conclusion. In such cases, we suggest that the necessary additional information could be obtained by HPLC or GC of the oxidised UCM as suggested by Revill *et al.* (1992).

Since the reliability of pattern recognition methods is undermined by extraneous variability, it is important to ensure consistency of protocols within a given investigation. Thus, the GCs of field and potential source samples should be produced by the same laboratory using the same protocol and, ideally, the same instrument.

6.3 Data Processing and Databases

6.3.1 Identification of key characteristics

Effective computer-based pattern matching depends upon the input data containing the key discriminating features, and preferably little else. If data are to be draw from the entire GC trace then it is desirable to reduce the number of inputs to just those peaks which are important discriminators. These will depend upon the range of oil types that are to be classified by the system. The EUROCRUDE system for the identification of crude oils uses 15 biomarker peaks (Grigson and Baron, 1993; Sinclair and Grigson, 1996), although the database that was constructed for its development contained data for 56 peaks. The system developed by Wigger and Torkelson (1997) for the identification of petroleum hydrocarbons initially used 71 GC peaks, but this was later extended to 89 peaks. Long *et al.* (1991) used 48 GC peaks to classify jet fuels and Welsh *et al.* (1996) used 46 HPLC peaks to classify a pharmaceutical product, after first testing their systems using 22, 46 and 899 (i.e. the whole HPLC output) input peaks. Lavine *et al.* (1998) classified jet fuels using an optimal 20-peak feature vector, which they derived using a genetic algorithm.

If the Agency wishes to use computer-based pattern recognition to identify the type of oil responsible for an oil pollution incident, then it will be necessary to determine the set of GC peaks that will enable the system to discriminate between all of the specified oil types. The optimum membership of this set could be derived using a genetic algorithm, neural network or information theory. The need to do this does not arise with the proposed method of identifying the source, because this uses all of the peaks within small sections of the GC, and either each section is sufficiently characteristic of the true source to provide a reliable match or it is not. Nevertheless, it is necessary in both cases to reliably match corresponding peaks in the GC traces that are to be compared. Unfortunately, retention times from one GC to another are not normally directly comparable owing to retention time shift from test to test, so it is necessary to rescale retention times before attempting to match corresponding peaks. This can be done by uniformly rescaling the time axis between the peaks of two or more known components of the sample (or introduced internal standards).

6.3.2 Effects of weathering

Weathering causes preferential outwash and evaporation of the lighter, more volatile, components of the spilled oil, and so has its greatest impact on the GC peaks with low retention times. This causes an apparent mismatch between field samples and their original oil, especially in the lower end of the GC. This can be largely overcome by appropriate rescaling of the intensity axis, but how is this best achieved? If the response axis (i.e. y-axis) of the GC is to be rescaled over the whole of its range then this needs to be done as a function of retention time, so as to allow for the differential effect of weathering with respect to retention time. Our exploratory investigation indicated that a satisfactory rescaling could be achieved using a simple function of retention time, but further investigation of this is clearly necessary. An alternative approach would be to develop a model of the weathering process, as was done by Wigger and Torkelson (1997). However, where the GCs are to be compared in terms of a series of corresponding small sections of the overall trace, then the time component of the rescaling becomes far less significant within individual sections. In this case, a linear rescaling

of the form $y' = ay + bx + c$ would be quite adequate. The values of the constants a, b and c could be optimised for each particular section and field-to-source comparison.

6.3.3 Construction of a database

The development of a computer-based system to classify the type of oil responsible for an oil spill would require a database of good quality GC traces that are representative of the whole range of oil types that are to be classified. One problem to be addressed in developing this database is that of consistency. There is a need for a high degree of consistency in the way the GC traces are produced, but there is also a need for them to be representative of the GCs that will be presented to the system for classification by its various users. Our current thinking on this issue is that the GC traces should be derived from the whole range of laboratories that are likely to want to use the system, using exactly the same analytical protocol.

There is also a need to develop a database of case studies to provide the necessary data on which to test the performance of the source identification system. Each case would have to include GC traces of the field samples, the true source and a number of potential sources. Furthermore, all of the data for an individual case must have been produced by the same laboratory under the same protocol, and preferably using the same instrument.

6.3.4 Data quality assurance

It is important that all data used, whether to compile the database of standard oil types or to identify the source of a particular oil spill, be produced by fault free equipment. Whilst this would normally be achieved by internal laboratory quality assurance procedures, it is possible to further enhance these procedures through the use of pattern recognition systems. Elling *et al.* (1997) developed a hybrid artificial intelligence tool for assessing GC data. The system combines a rule-based system with pattern recognition based on neural networks to identify instrument faults/malfunctions from patterns in the GC data. Although this is not vital to the development of a system for the identification of the source of oil spills, the Agency should consider the possibility of developing such as system in the long term. If it decides to proceed, we recommend that the expert systems component of the system be based upon a Bayesian belief network (BBN), not a rule-based system, because the problem involves inherent uncertainty and rule-based systems cannot reason properly under conditions of uncertainty. Work currently being carried out as part of National R&D Project E1-056 'Development of AI Systems for the Diagnosis of River Quality' is based on the use of BBNs and could be relevant to this matter.

6.4 Choice of Pattern Recognition Methods

It is well known that the relative merits of different methods of classification vary between data sets. Published results of projects that have used neural networks and standard statistical methods to analyse GC data indicate that neural networks may be marginally better than standard statistical methods, but this is offset by the likelihood that some of these networks may have been overfitted to the data (see section 4.2.2). It is not therefore possible at this stage to say that one technique is better than the other.

The choice of method also depends upon the particular problem. When attempting to identify the type of oil that is contaminating a river, it will be possible to use a large database of standard types, so it should be possible to train a neural network without the danger of overfitting the data. Under these circumstances, the best course of action would be to test supervised-learning networks (e.g. backpropagation MLP) and unsupervised-learning networks (e.g. SOMF) against a range of statistical methods. The work carried out in National R&D Projects E1/i621 'Applications of AI in River Quality Surveys and E1-056 'Development of AI Systems for the Diagnosis of River Quality' may be of value with respect to the choice of techniques.

On the other hand, when attempting to match field samples with samples from potential sources, the amount of data is very limited and the use of a neural network would be inappropriate. There are, however, several statistical methods that could be used, and the best approach would be to test several methods using a set of case studies.

6.5 Development of a Computer-based System

Our view, as stated earlier, is that the development of computer-based system for identifying the source of an oil spill on inland waters would best be developed using a two-stage approach. First, to classify the oil type, then to identify its source.

6.5.1 Oil type classifier

This system would be designed to read the input vector of the *n* key GC peaks (principally the n-alkanes, pristane, phytane etc.) of a field sample and then classify it to one of the standard oil types. The input vector could also incorporate some of the features that are currently used in GC analysis, like the pristane/phytane ratio, and individual inputs could be weighted according to their relative importance as indicators. Provided that the field sample has only been weathered to a moderate degree it should be possible to classify it directly, without making allowance for the effect of weathering, except perhaps if it is a jet fuel. If it is severely weathered it will be necessary to account for the effect of weathering within the pattern matching process. This can either be done by developing models that predict the effects of varying degrees of weathering on GCs, or by designing the system to identify the oil directly, in which case the database used for its development would have to include representative examples of weathered standard oils. The former approach was adopted by Wigger and Torkelson (1997), who developed an algorithm to model the evaporation component of the weathering of gasoline. This could be further developed (via controlled weathering experiments) to express the effects of weathering as well-defined mathematical transformations of GC data, thus facilitating 'intelligent' guidance for a system designed to find matches between GCs irrespective of the degree of weathering. However, both of these approaches would require data on the effects of weathering on the full range of standard oils.

Clearly, a preliminary investigation needs to be carried out to determine: a) the magnitude of the pattern matching problem created by a typically weathered field sample; and b) the most cost-effective way of overcoming the problem. In the short term the answer may be to continue using visual methods of identifying the type of oil responsible for the spill.

6.5.2 Source detector

All the current methods of source detection rely on the use of well-resolved specific compounds (e.g. the n-alkanes, pristane and phytane) or ratios of these (e.g. pristane/phytane). Even when matching chromatograms by eye, it is natural to concentrate first on the major peaks. Although this is generally sufficient for identification of a particular type of oil, it is often insufficient when trying to discriminate between several potential sources of the same type. However, there is a wealth of additional features in the minor peaks that lie in the sections of the GC between the major peaks. This study has shown that the patterns of peaks within some of these sections of the field samples may be unique to the true source. Thus we recommend that a pattern matching system for source identification be based primarily on these characteristic, as outlined in section 5.2.5. This approach, of matching field sample to source samples in sections of the GC and then combining the results, was also used by the developers of EUROCRUDE. However, the design of the system will have to be different to that of EUROCRUDE, which was based on a very limited set of main peaks and a large database of potential sources, whereas the proposed system will be based on all of the minor peaks and a very limited data set. It is likely that some evidence of the true source will also be provided by the pattern of major peaks over the whole GC (i.e. the pattern that we suggest be used to identify the oil type), but we anticipate that this will be of secondary importance. Nevertheless, the system should be designed to combine this evidence with that derived from the minor peaks when drawing its overall conclusion. This would best be achieved using Bayesian methods. It should be noted that the development, testing and validation of the system will require the construction of a database of case studies, as mentioned earlier in section 6.3.3.

6.6 Utility of the System

An important aim of this study was to investigate whether advanced computer-based techniques of pattern comparison might increase the reliability and defensibility of oil identification relative to visual methods. Our review of the scientific literature produced some mixed opinions on the merits of computer-based interpretation relative to interpretation by human experts. Studies by Welsh *et al.* (1996) and Rowe *et al.* (1994) claimed that their computer systems performed better than humans. However, other studies (Duller *et al.*, 1996; Wigger *et al.*, 1998) concluded that their systems did not out-perform human experts or de-skill the interpretation of GC data, but that further improvements were possible. All concluded that their systems were valuable tools for the interpretation of chromatographic/spectrographic data. The main advantages quoted were the objective nature and processing speed of computer-based systems. Our own study indicates a possible further advantage - the ability to analyse and identify patterns in the minute detail of GCs - a process that would be difficult and time consuming by eye. Perhaps the most significant comment on the utility of computer-based pattern recognition systems is the statement by ESR in New Zealand, that EUROCRUDE "passed with flying colours" when they used it to prosecute a case of oil pollution in Lyttleton Harbour (See http://www.esr.cri.nz/services/analytical/success-stories.html).

We conclude that the development of a computer-based pattern recognition system to identify the source of an oil spill on an inland water would certainly improve the reliability and defensibility of evidence given in court. It would not only be seen to provide objective evidence, but also it would readily identify the unique characteristics that exist between field and source samples, albeit in very small sections of their fingerprints.

6.7 Recommendations for Future Research

6.7.1 Priorities

There are a number of options available to the Environment Agency as regards future research in this area. Clearly, it would be possible to embark on the development of a comprehensive computer-based source identification system that incorporated data quality assurance, oil type identification and source identification. However, the Agency might prefer to take a more cautious staged approach in which the most beneficial and innovative components are proven first. In this case our recommended order of development is as follows.

1. Carry out a detailed feasibility study of the proposed source identification system based upon patterns of minor peaks in sections of the GC trace. If shown to be feasible, develop and field test a user-friendly system. Note that this approach assumes that the identification of the oil type responsible for the spill will be continue, for the time being, to be done by eye.
2. If the source identification system proves its worth, then proceed with the development and testing of a system to automatically identify the type of refined oil responsible for the spill. This exercise would involve experimentation to determine the best procedures and techniques to use.
3. If this proves successful, consider developing a system for the quality assurance of GC data and the diagnosis of faults in analytical instruments. This is not vital to the overall system, but would add another level of quality assurance to its results.

Alternatively, if the Agency wishes to develop a comprehensive system in a single project, it would have to combine these three components into one integrated project.

The following sections provide details of the research programmes necessary for the three stage approach outlined above.

6.7.2 Development of a source identification system

This system is the most essential component of the overall system and could operate as a stand alone system, provided the identification of oil type continued to be done by eye. It is therefore the most beneficial system to develop first, but it is also the most innovative and hence risky. Our recommended programme of work is outlined below.

- Construct a database of case studies, consisting of at least 20 oil spill events covering a wide range of oil types, and including GCs of field samples and samples from several potential sources derived using consistent protocols.
- Construct a database of several GC replicate test results for each of the standard oil types. This is required to enable experimental variability under standard conditions to be quantified. If the Agency does not possess such data, then it will be necessary for the contractor or the Agency to carry out a series of tests to produce them.
- Develop and test software (based on the method outlined in sections 5.2.5, 5.2.6 and 6.5.2) to identify the true source by comparing the GCs of the field samples and potential sources.

- Investigate the magnitude of differences between GCs that occur due to experimental variability, using the database of replicate test results.
- Develop software to combine the conclusions drawn from the various sections of the GCs to produce an overall conclusion, with an associated measure of certainty.
- Develop software to produce graphical displays illustrating the match between the field samples and the true source, and also the mismatch between the field samples and the other potential sources.
- Evaluate the ability of these prototype systems to provide the basis of a robust operational system.
- If the outcome of this evaluation is positive, develop a user-friendly source identification and data presentation system for field testing in the Agency's laboratories and, if possible, test the value of its outputs by using them as evidence in specific court cases.
- If this proves satisfactory, fine tune the system based on feedback from the field trial; and deliver the final operational system.

6.7.3 Development of a system to identify standard oil types

This is the second most important component of the overall system, but it would have limited value as a stand-alone system. Its main value would be realised when integrated with the source identification system. It could, however, be used simply to take the tedium out of routine classification of GCs traces into oil types. The results of our literature review have shown that the development of this system is certainly viable, but they did not clearly indicate the best procedures or techniques to use. Therefore the project will, of necessity, involve an investigation to determine the best way to proceed. Our recommended programme of work is outlined below.

- Construct a database of at least 300 GCs (but preferably over 500) covering the whole range of standard oils. It is assumed that these data will be available from archives, but if not they will have to be produced by the contractor or the Agency.
- Construct a database consisting of as many GCs as possible of weathered standard oils, covering varying degrees of weathering and as wide a range of oils as possible. If the Agency does not possess such data, then it will be necessary for the contractor or the Agency to carry out a series of tests to produce them using a range of standard oils and various degrees of controlled weathering.
- Develop several oil type classifiers using various statistical and neural network pattern matching techniques and the database of unweathered standard oils. Test their ability to correctly classify a) unweathered standard oils; and b) weathered field samples. The results of (a) will enable the accuracy of the different pattern matching techniques to be compared, and the results of (b) will enable the impact of weathering on success rates to be assessed.
- If weathering is found to have a noticeable effect on success rates, then it will be necessary to explore ways of accommodating its effect within the pattern matching process. This could be achieved by:
  a) training and testing neural network classifiers using training and testing data sets based upon the combined unweathered and weathered databases of standard oils; or

b) developing a mathematical model of weathering based on the database of weathered standard oils, and then using it to 'unweather' the field samples prior to classifying them using the previously derived 'unweathered' classifiers.

The neural network approach would require good representation of weathered samples in the database to ensure success, but a satisfactory mathematical model could probably be achieved using fewer weathered samples, especially if produced by a laboratory programme of controlled weathering experiments.

- Assess the performance of the various classifiers and methods of dealing with weathering, and select the most promising approach. In the light of the results of the testing and assessment exercises, make any necessary improvements to the classifier.
- Integrate the oil type classifier with the source identification software to produce a user-friendly integrated package for field testing under operational conditions.
- Field test and fine tune the system based upon feedback from users, and deliver the final operational system.

6.7.4 Development of a GC quality assurance system

This system would add an extra degree of quality assurance to the conclusions drawn from the integrated oil type classifier and source identification system. Our recommended programme of work is outlined below.

- Construct a database of faulty GCs that were produced by instruments that were found to be malfunctioning. This must include data on the cause(s) of the problem.
- Develop an integrated neural network / expert system to examine GC traces and diagnose any malfunction of the instruments that produced them. The paper by Elling *et al.* (1997) provides a starting point for the development of this system, except that we strongly recommend the use of a Bayesian belief network approach to the development of the expert system component, instead of a rule-based approach.
- Once this system has been tested and proven it should be developed as a user-friendly GC data validation system, and not integrated with the main system.

7. CONCLUSION

A comprehensive literature search has shown that statistical and neural network techniques of pattern recognition have been successfully used to identify standard refined oils and the source of crude oil spills affecting coastal areas. However, no evidence has been found of the use of pattern recognition techniques to identify the source of an oil spill affecting inland waters. It has been shown that oil spills on inland waters present a different kind of problem to that addressed by earlier studies, owing to the inevitable sparcity of data on the potential sources. It is concluded that in this case a two-stage approach to the problem is required, the first to identify the type of refined oil responsible for the spill, and the second to identify the true source from all potential sources. Potential solutions to the first task are well known, since it presents a similar problem to that addressed by many earlier studies. A potential solution to the second task has been found via an exploratory investigation based upon a visual pattern recognition exercise involving a specific case study. It is concluded that the use of computer-based pattern recognition techniques would improve the reliability and defensibility of evidence given in court, and recommendations are made for a phased programme of research leading to the development of a comprehensive source identification system.

8. ACKNOWLEDGEMENTS

The authors wish to thank David Britnell (Project Manager for the Environment Agency) and David Gazzard (Senior Scientist at the Environment Agency's Laboratory at Fobney) for their help in defining the problem and providing data for the exploratory investigation.

9. KEY REFERENCES

9.1 Scientific Papers

Duller A. W. G., Hatton R. S., Wells B. T. and Barwise A. (1996) Application of neural networks in crude oil fluorescence fingerprinting. In *Artificial Intelligence in the Petroleum Industry 2, Chapter 6,* (eds. Braunschweig B. and Bremdal B. A.), 139-158. Editions Technip.

Elling J. W., Lahiri S., Luck J. P., Roberts R. S., Hruska S. I., Adair K. L., Levis A. P., Timpany R. G., and Robinson J. J. (1997) Hybrid artificial intelligence tools for assessing GC data. Analytical Chemistry, **69** (13), 409A- 415A

Grigson S. J. W. and Baron G. R. (1993) The European approach to the source identification of oil spills: a study of its specificity and reliability. Oil Spill Conference, Tampa, Florida.

Grigson S. J. W. and Baron G. R. (1995) The European Crude Oil Identification System. German Journal of Hydrography, Vorträge des 5. Internationalen Wissen schaftlichen Symposiums, Hamburg, 127-136.

Hendrick M. A. and Jadamec J. R. (1991) Evaluating the relative performance of ASTM methods in the laboratory and the field. In: *Monitoring water in the 1990's: meeting new challenges*, ASTM Special Publication, 567-577.

Koussiafes P. and Bertsch W. (1993) Profile matching for the analysis of accelerants in suspect arson cases. J. Chromatogr. Sci., **31** (4), 137-144.

Lavine B. K., Moores A. J., Mayfield H. T. and Faruque A. (1998) Fuel spill identification by gas chromatography - genetic algorithms/pattern recognition techniques. Analytical Letters **31** (15), 2805-2822.

Long J. R., Mayfield H. T., Henley M. V. and Kromann P. R. (1991) Pattern recognition of jet fuel chromatographic data by artificial neural networks with back-propagation of error. Analytical Chemistry, **63**, 1256-1261.

Mason J. P., Kirk I., Windsor C. G., Tipler A., Spragg R. A. and Rendle M. (1992) A novel algorithm for chromatogram matching in qualitative analysis. Journal of High Resolution Chromatography, **15**, 539-547.

Revill A. T., Carr M. R. and Rowland S. J. (1992) Use of oxidative degradation followed by capillary gas chromatography-mass spectrometry and multi-dimensional scaling analysis to fingerprint unresolved complex mixtures of hydrocarbons. Journal of Chromatography, **589**, 281-286.

Rowe R. C., Mulley V. J., Hughes J. C., Nabney I. T. and Debenham R. M. (1994) Neural networks for chromatographic peak classification - a preliminary study. LC-GC International **7**, 36-42.

Sinclair J. W. and Grigson S. J. W. (1996) Oil spill fingerprinting - the practical benefits to the operator. Proc. of Conf. on Health, Safety and Environment, New Orleans, Louisiana, 811-819.

Stout S. A., Uhler A. D., Naymik T. G. and McCarthy K. J. (1998) Environmental forensics: unraveling site liability. Environmental Science and Technology / News, June, 260-264.

Walley W. J., Fontama V. N. and Martin R. W. (1998) Applications of artificial intelligence in river quality surveys. R&D Technical Report E52. Environment Agency, Bristol.

Wang Z. and Fingas M. (1997) Developments in the analysis of petroleum hydrocarbons in oils, petroleum products and oil-spill-related environmental samples by gas chromatography (A Review). Journal of Chromatography A, 774, 51-78.

Welsh W. J., Lin W., Tersigni S. H., Collantes E., Duta R., Carey M. S., Zielinski W. L., Brower J., Spencer J. A. and Layloff T. P. (1996) Pharmaceutical fingerprinting: evaluation of neural networks and chemometric techniques for distinguishing among same-product manufacturers. Analytical Chemistry, 68 (19), 3473- 3482.

Wigger J. W., Beckmann D. D., Torkelson B. E. and Narang A. X. (1998) Petroleum hydrocarbon fingerprinting quantitative interpretation: development and case study for use in environmental forensic investigations. National Groundwater Association / American Petroleum Institute - Proc. of Conference on Petroleum Hydrocarbons in Groundwater, Houston TX. (Also on the Web at http://www.elmengineering.com/)

Wigger J. W. and Torkelson B. E. (1997) Petroleum Hydrocarbon Fingerprinting - Numerical Interpretation Developments. Proc. of the 4th Annual International Petroleum Environmental Conference. San Antonio, Texas. (Also on the Web at http://www.elmengineering.com/)

9.2 Commercial Literature

Pirouette - Technical Publications by InfoMetrix (http://www.infometrix.com) relating to their pattern recognition software package, include:

1. Applications Overview: Chemometrics in Environmental Science.
2. Application Overview: Chemometrics in Chromatography.
3. Technical Note: Description of Pirouette Algorithms.

MatchFinder - Information is published on the Web by AEA Technology plc relating to their pattern recognition software package (http://www.aeat.co.uk/pes/software/match.html).

Advanced Chemical Fingerprinting: A critical set of tools for companies facing liability claims. Information published on the Web by Boehm P. D., Douglas G. S. and Brown J. S. of Arthur D Little, Inc., Environmental, Health and Safety Consulting (http://www.arthurdlittle.com).

APPENDIX A

Results of Literature Search
-
Lists of Papers not included in Key References

Initial Search Results (mainly neural network references)

Rivera, S.L. & Klein, E.J. "Automatic classification of chromatographic peaks" Proc. of the 1997 American Control Conference 1997 5:3262-3266

Abstract: *An intelligent algorithm was developed to automatically categorize chromatographic peaks resulting from the separation of protein mixtures using ion exchange chromatography. A vector quantizing neural network (VQN) was trained and used to classify peaks into six distinct categories based on peak geometry: Gaussian, fronted, tailed, leading shoulder, trailing shoulder, and overlapping. A preprocessing algorithm consisting of noise filtering. vector normalization, and cubic spline interpolation was developed to map peaks to identically sized vectors before introducing them to the VQN. Experimental data was used for training and testing. The VQN correctly classified 90% of the test peaks.*

Rowe, R.C.; Mulley, V.J.; Hughes, J.C.; Nabney, I.T.; Debenham, R.M. "Neural networks for chromatographic peak classification - a preliminary study" LC-GC International 1994 7:36-42

Abstract: *A multi-layer perception neutral network was trained for peak-shape classification in chromatography. Three classes of peak profiles were considered as suitable data sets: "good"; "tailing"; and "unresolved" (details given). The neural network architecture, to determine the most suitable network shapes and values for learning parameters, is described as well as cross-validation experiments. Finally the trained network was compared with a human expert in the classification of 396 individual peak profiles. Both exhibited a success rate of 85%, however the neural network performed the task in 5.6 s whereas the human took 8 h. The neural network showed complete objectivity. Results are discussed.*

Hendrick,MS & Jadamec,JR "Evaluating the relative performance of ASTM methods in the laboratory and the field." ASTM Special Technical Publication 1991 No.1102. pp.567-574

Abstract: *In order to establish legal responsibility for an oil spill and recover clean-up costs, the U.S. Coast Guard Oil Identification Laboratory compares spilled and suspected source oil samples using a multi-method approach based primarily on ASTM standards. This paper describes a computerized evaluation of whether improvements in technology and subsequent method development have improved the relative performance of the analytical methods. Artificial intelligence software capable of analyzing data bases containing qualitative as well as quantitative information was employed. Results of this on-going laboratory study indicate that the revised Gas Chromatography method performs better, relative to the Fluorescence, Infrared, High Performance Liquid Chromatographic and Thin Layer Chromatographic methods. A field study was conducted during the deployment of the Coast Guard Research and Development Center's Mobile Lab to Alaska where it was used to assist in assessing the environmental impact of the spilled EXXON VALDEZ oil. Over twelve hundred samples were analyzed and 'fingerprinted' to determine if they were related to the spilled cargo oil. The rapid turn-around times required and a variety of operational, sampling-and weathering problems presented difficulties not normally encountered in normal laboratory operation. The approach that proved to be highly successful included screening and fingerprinting samples with fluorescence techniques, including ASTM standard method D3650 for emission spectra as well as synchronous scanning techniques. Samples found to be similar to the EXXON VALDEZ cargo were analyzed by GC/MS techniques for confirmation. This approach met the rapid turn around times requested and demonstrated that a multi-method approach is the most efficient and reliable way to positively match a spilled oil to a common source.*

Beebe, K.R.; Blaser, W.W.; Bredeweg, R.A.; Chauvel, J.P. Jr; Harner, R.S.; LaPack, M.; Leugers, A.; Martin, D.P.; Wright, L.G.; Yalvac, E.D. "Process analytical chemistry." Anal. Chem., 1993, Vol.65, No.12, pp.199R-216R

Abstract: *A review is presented, with 507 references, of the literature published between 1987 and 1992 on analytical techniques used in process control. Techniques included are chromatography, optical spectroscopy, fibre optics, MS, chemometics, artificial neural networks and FIA.*

Coenegracht, P.M.J.; Metting, H.J.; VanLoo, E.M.; Snoeijer, G.J.; Doornbos, D.A. "Peak tracking with a neural network for spectral recognition." J. Chromatogr., 1993, Vol.631, No.1-2, pp.145-160

Abstract: The tracking procedure uses UV spectra (collected by a multi-channel diode-array spectrometer) and peak areas measured atone wavelength. The spectra of all components were measured in one chromatogram, and after normalization and multiplication by the corresponding peak areas the data were used to train a neural network with back-propagation consisting of an input, hidden and output layer. All calculations and data processing were carried out on a PC with a mathematical co-processor. The technique is illustrated by the successful peak recognition of the separation of a group of eight sulfonamides with considerable peak overlapping.

Liu, Y.; Upadhyaya, B.R.; Naghedolfeizi, M. "Chemometric data analysis using artificial neural networks." Appl. Spectrosc., 1993, Vol.47, No.1, pp.12-23

Abstract: *Hybrid signal reprocessing and artificial neural network paradigms have been applied to online composition analysis of chemical samples from chemometric data and the performance of the methodology tested with the use of near-IR and Raman spectra of industrial and laboratory samples containing mixtures of aromatic and aliphatic hydrocarbons. The sensitivity of composition estimation as a function of spectral errors, spectral pre-processing, choice of parameter vector, the optimal architecture of multilayer neural networks and guidelines required to achieve these objectives were studied. The neural network method can be easily applied to other spectroscopic data applications such as lubrication oil analysis, effluent gas analysis, water chemistry, etc.*

Goodacre, R.; Kell, D.B.; Bianchi, G. "Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks." J. Sci. Food Agric., 1993, Vol.63, No.3, pp.297-307

Abstract: *Rapid assessment of the adulteration of extra-virgin olive oils with other seed oils was achieved by a combination of Curie-point pyrolysis MS with multivariate data analysis using artificial neural networks. Samples included a variety of representative cultivars, crushing protocols and storage regimes. Pyrolysis MS was with use of a Horizon Instruments PyMS-200X and data was collected over m/e 51-200. The method was rapid (sample time 2 min) and could accurately assess contamination of virgin olive oils adulterated with 50-500 ml of corn, peanut, soya, sunflower or rectified (sansa) olive oil per 1 of mixed oil.*

Andrews, J.M. & Lieberman, S.H. "Neural network approach to qualitative identification of fuels and oils from laser-induced fluorescence spectra." Anal. Chim. Acta, 1994, Vol.285, No.1, pp.237-246

Abstract: *Seven classes of petroleum hydrocarbon-based fuels and oils (listed) can be identified from their fluorescence emission spectra (FES) using a series of software implemented, three-layer, back-propagation neural networks. The N2 laser-induced fluorescence spectra of multiple samples of each class are collected through an optical fibre and incorporated into a multi-channel detection system. Thirty-six spectral examples of the seven fuels and oil classes are partitioned into seven separate paired combinations of training and test spectra. Each combination employs 29 spectra to independently train a network and the spectra of the seven remaining samples are used to test the trained network's ability to make generalized classifications. Networks trained with data sampled directly from the normalised FES identify 96% of the test spectra accurately. An additional series of networks is trained and tested using the same spectra but with principal component analysis (PCA) employed as a pre-processor. Networks trained with PCA processed spectral data achieve a lower performance and identify only 90% of the test spectra successfully.*

Davies, A.M.C. "Classification by artificial neural networks." Spectroscopy Europe, 1994, Vol.6, No.5, pp.27-29

Abstract: *Three applications of artificial neural networks to classification, which utilize data from chromatography, NIR spectroscopy and FT-Raman spectroscopy to the problems of olive oil identification, plastic waste classification and wood hardness, respectively, are described.*

Zupan, J.; Novic, M.; Li, X.Z.; Gasteiger, J. "Classification of multicomponent analytical data of olive oils using different neural networks." Analytica Chimica Acta, 1994, Vol.292, No.3, pp.219- 234

Abstract: Samples of 572 olive oils from nine regions of Italy were analysed for eight fatty acids (viz, palmitic, palmitoleic, stearic, oleic, linoleic, arachidic, linolenic and eicosenoic acids). Because of the large differences in concentrations of these acids, the results were scaled from 0-100 with respect to the range between the lowest and highest concentration of each. These data were then used in the classification study. Part of the set of samples from each region was used as a training set and the remainder as a test set. Two neural networks were compared for their ability to classify the oils correctly by region. Kohonen learning gave better results than back-propagation of errors. With this method, only 16 of the 322 samples in the test sets were wrongly assigned, mainly at the regional boundaries. The weightings used in the Kohonen learning also gave information about the significance of each fatty acid in the assignments.

Goodacre, R.; Kell, D.B.; Bianchi, G. "Food adulteration exposed by neural networks." Analysis Europa, 1995, No.5, pp.35-37

Abstract: The use of neural networks in combination with pyrolysis-MS for detecting adulteration of virgin olive oil is discussed. The neural networks were trained using the standard back- propagation method and the effectiveness of the training was expressed as the root-mean-square error; a value of 0.1% was attained. The method correctly assessed each oil and it is hoped that the level of contamination may also be assessed in the future. As any biological material can be pyrolysed, the method may be applied to the contamination of any food material.

Husain, S.; Devi, K.S.; Krishna, D.; Reddy, P.J. "Characterization and identification of edible oil blends and prediction of the composition by artificial neural networks - a case study." Chemometrics and Intelligent Laboratory Systems, 1996, Vol.35, No.1, pp.117-126

Abstract: Vegetable oils and oil blends were esterized and analysed by GC with FID (experimental details given). The results were subjected to chemometric analysis using artificial neural networks (ANN) and multiple linear regression (MLR). The training set consisted of GC peaks for palmitic, stearic, oleic, linoleic and linolenic acids for pure groundnut oil and 90% blends with other edible oils. The test sets were GC data for 50%, 60%, 70% and 80% blends. The ANN model in each case correctly identified whether the sample was a pure oil or a binary blend. A MLR model based on the fatty acid ratios in binary oil blends had regression coefficients of >=0.967 and F-test values were significant at the 5% level for all oil blends tested. The model could be used to predict the composition of an oil blend.

Elling, J.W.; Klatt, L.N.; Mniszewski, S.M. "Automated chromatography data interpretation using an expert system to integrate standard and pattern recognition data processing techniques." US Dept. of Energy paper available on-line at:
http://thermal.esa.lanl.gov/DataInterpretation/automatedprocessing/pattern_recognition/pcr/acdiues.html

Song, X.H. & Hopke, P.K. "Kohonen neural network as a pattern-recognition method based on the weight interpretation." Analytica Chimica Acta, 1996, Vol.334, No.1-2, pp.57-66

Abstract: The self-organizing Kohonen neural network (K-NN) is a useful tool for pattern recognition and predictions can be made for unknown objects based on the Kohonen map obtained from a training set. The membership of new objects hitting empty neurons that were not activated by any training set objects has been achieved with the K-nearest neighbour technique (discussed), but some information about the correct neighbour relationships between the object vectors was lost during the projection into a low-dimensional subspace for the K- NN. An alternative procedure, based on the weight interpretation (K- WI), permitted the K-NN to be used in a supervised way. The membership of samples that hit empty neurons during the prediction process was determined to be the same as that of the nearest active neuron in terms of a distance measure from the trained weight vectors. This procedure was applied to the "Italian olive oil" data set. The K-WI procedure gave better prediction results than the K-NN. The LVQ method (discussed) gave classification results that were similar to and as satisfactory as those given by K-WI, although the latter was the easier to use.

Eghbaldar, A.; Forrest, T.P.; CabrolBass, D.; Cambon, A.; Guigonis, J.M. "Identification of structural features from mass spectrometry using a neural network approach: application to trimethylsilyl derivatives used for medical diagnosis." Journal of Chemical Information and Computer Science, 1996, Vol.36, No.4, pp.637-643

Abstract: *Organic acids were extracted from urine with ether. The extracts were dried with anhydrous Na2SO4 under N2. TMS derivatives were prepared and applied to a DB1 or DB5 capillary column (30 m * 0.25 mm i.d., further details not given). Mass spectra were recorded and an artificial neural network (ANN) was applied to identify specific structural features. ANN combined with other taxonomic methods was used for structural determination. The input vector was composed of the intensities of peaks with 40<=m/z<=220. The output vector was generated from the known structure. The response ratio and the quality of responses were high for all investigated structural features. The method was used for rapid acidemias diagnosis.*

Papazova D. and Pavlova A. "Development of a Simple Gas Chromatographic Method for Differentiation of Spilled Oils." Journal of Chromatographic Science, 1999, 37 (1), 1-4.

Abstract. An approach for the fast, preliminary identification and differentiation of fresh oil spills is proposed.Capillary gas chromatography with flame ionization detection for the determination of n-alkane and isoprenoid distribution in oil spill samples is applied. An internal standard method is used for the quantitation of the selected compounds. Five characteristic parameters are checked for adequate presentation. n-Alkanes and isoprenoids are chosen as the most suitable structures for the identification and differentiation of fresh oil spills. In many cases, this information is sufficient to eliminate most of the oils as potential sources of the pollution.

References from Infometrix (mainly pattern recognition)

Musumarra, G.; Scarlata, G.; Romano, G.; Cappello, G.; Clementi, S. and Giulietti, G. "Qualitative Organic Analysis. Part 2. Identification of Drugs by Principal Components Analysis of Standardized TLC Data in Four Eluent Systems and of Retention Indices on SE 30." J. Anal. Toxicology (1987) 11 (Jul./Aug.): 154-163.

Engman, H.; Mayfield, H.T.; Mar, T. and Bertsch, W. "Classification of bacteria by pyrolysis-capillary column gas chromatography-mass spectrometry and pattern recognition." J. Anal. Appl. Pyrolysis (1984) 6 (2): 137-156.

Butler, W.R.; Jost, K.C. and Kilburn, J.O. "Identification of mycobacteria by high performance liquid chromatography." J. Clin. Microbiol. (1991) 29 (11): 2468-2472.

Kowalski, B.R. "Measurement analysis by pattern recognition" Anal. Chem. (1975) 47:1152A-

Marshall, R.J.; Turner, R.; Yu, H. and Cooper, E.H. "Cluster analysis of chromatographic profiles of urine proteins." J. Chromatogr. (1984) 297: 235-244.

Pino, J.A.; McMurry, J.E.; Jurs, P.C. and Lavine, B.K. "Application of pyrolysis/gas chromatography / pattern recognition to the detection of cystic fibrosis heterozygotes." Anal. Chem. (1985) 57 (1): 295-302.

Moret, I.; Scarponi, G. and Cescon, P. "Aroma components as discriminating parameters in the chemometric classification of Venetian white wines." J. Sci. Food Agric. (1984) 35 (9): 1004-1011.

Moret, I.; Scarponi, G.; Capodaglio, G. and Cescon, P. "Characterization Soave wine by determining the aromatic composition and applying the SIMCA chemometric method." Riv. Vitic. Enol. (1985) 38 (4): 254-262.

Stenroos, L.E. and Siebert, K.J. "Application of pattern-recognition techniques to the essential oil of hops." J. Am. Soc. Brew. Chem. (1984) 42 (2): 54-61.

Van Rooyen, P.C.; Marais, J. and Ellis, L.P. "Multivariate analysis of fermentation flavor profiles of selected South African white wines." Dev. Food Sci. (1985) 10 (Prog. Flavour Res.): 359-385.

Saxberg, B.E.H.; Duewer, D.L.; Booker, J.L. and Kowalski, B.R. "Pattern recognition and blind assay techniques applied to forensic separation of whiskies." Anal. Chim. Acta (1978) 103: 201-212.

Zumberge, J.E. "Prediction of source rock characteristics based on terpane biomarkers in crude oils: A multivariate statistical approach." Geochim. Cosmochim. Acta (1987) 51 (6): 1625-1637.

Dunn, W.J.; Stalling, D.L.; Schwartz, T.R.; Hogan, J.W.; Petty, J.D.; Johansson, E. and Wold, S. "Pattern recognition for classification and determination of polychlorinated biphenyls in environmental samples." Anal.Chem. (1984) 56 (8): 1308-1313.

Onuska, F.I.; Mudroch, A. and Davies, S. "Application of chemometrics in homolog-specific analysis of PCBs." HRC & CC (1985) 8: 747-754.

Breen, J.J. and Robinson, P.E., Eds., Environmental Applications of Chemometrics ACS Symposium Series (1985) 292: 286pp.

Chien, M. "Analysis of complex mixtures by gas chromatography/mass spectrometry using a pattern recognition method." Anal. Chem. (1985) 57 (1): 348-352.

Isaszegi-Vass, I.; Fuhrmann, G.; Horvath, C.; Pungor, E. and Veress, G.E. "Application of pattern recognition in chromatography." Anal. Chem. Symp. Ser. (1984) 18 (Mod. Trends Anal. Chem., Pt. B): 109-124.

Smith, A.B.,.; Belcher, A.M.; Epple, G.; Jurs, P.C. and Lavine, B. "Computerized pattern recognition: a new technique for the analysis of chemical communication." Science (1985) 228 (4696): 175-177.

Stepanenko, V.E. "Group analysis and pattern recognition as a basis for chromatographic identification." Zh.Anal. Khim. (1985) 40 (5): 881-886.

Wenning, R. J. & Erickson, G. A. "Interpretation and analysis of complex environmental data using chemometric methods", Trends in Analytical Chemistry, 13:10, 1994, 446-457.

References from American Petroleum Institute:

Category Codes:

A - Analysis; C - Composition; I - Identification/Interpretation; S - Solubility

A	API. 1987. Manual of Sampling and Analytical Methods for Petroleum Hydrocarbons in Groundwater and Soil. API Publ. #4449. Amer. Petro. Inst., Wash. D.C.
A	API. 1987. Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. HESD Dept. Rpt. #214. Amer. Petro. Inst., Wash. D.C.
A	ASTM. 1983. Standard Test Method for Aromatic Hydrocarbons in Olefin-Free Gasolines by Silica Gel Adsorption. D 936-83. (Discontinued) Am. Soc. Testing Materials. Phil., PA
A	ASTM. 1986. Standard Specification for Automotive Gasoline. D 439-86 (discontinued, replaced by D 4814-93, see below). Am. Soc. Testing Materials. Phil., PA
A	ASTM. 1992. Standard Specification for Aviation Gasolines. D 910-92. Am. Soc. Testing Materials. Phil., PA (Previous standard was D 910-88a, published in 1988)
A	ASTM. 1992. Standard Specification for Fuel Oils. ASTM D 396-92. Amer. Soc. Testing Materials. Phil., PA (Previous standard was dated 1986, i.e., D 396-86)
A	ASTM. 1993. Standard Specification for Automotive Spark Ignition Fuel. ASTM D 4814-93. Am. Soc. Testing Materials. Phil., PA
A	ASTM. 1993. Standard Specification for Aviation Turbine Fuels. D 1655-93. Am. Soc. Testing Materials. Phil., PA (Previous standard was D 1655-88a, published in 1988)
A	ASTM. 1994. Analysis of Soils Contaminated with Petroleum Products. STP 1221. Am. Soc. Testing Materials. Phil., PA (9 peer reviewed papers on analysis and characterization)
A	Bell, A.C. 1994. Measurement of Trace Elements in Oil. Air & Waste Management Assoc. Annual Mtg, Cinn. OH Paper 94-MP6.01 (residual fuel oil)
A	Diehl, J. W., J. W. Finkbeiner, and F.P. DiSanzo. 1993. Determination of BTEX in Gasolines by Gas Chromatography/Deuterium Isotope Dilution Fourier Transform Infrared Spectroscopy. Analytical Chem. 65:2493-2496
A	Frick, C.S. 1987. Analytical Techniques for Soluble Component Analysis. IN: Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. Amer. Petro. Inst., Wash. D.C.
A	Johansen et al. 1983. Quantitative Analysis of Hydrocarbons by Structural Group Type in Gasoline and Distillates. I. Gas Chromatography. J. Chromatography, 256:393
A	Kanai, H. V. Inouye, R. Goo, R. Chow, L. Yazawa, and J. Maka. 1994. GC/MS Analysis of MTBE, ETBE and TAME in Gasolines. Anal. Chem. 66:924-927
A,	Klein, S.A. and D. Jenkins. 1981. The Quantitative and Qualitative Analysis of the Water Soluble Fraction of Jet Fuels. Water Research. 15:75–82.

A Levy, J.M and J.A. Yancey. 1986. Dual Capillary Gas Chromatographic Analysis of Alcohols and Methyl tert-/Butyl Ether in Gasolines. J. of High Resolution Chromatography & Chromatography Communications. 9:383–387.

A Lubeck, A. 1987. "Free" Product Analysis. IN: Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. HESD Dept. Rpt. #214. Amer. Petro. Inst., Wash. D.C.

A Mackay, D. et al. 1983. Testing of Crude Oils and Petroleum Products for Environmental Purposes. IN: Proceedings, 1983 Oil Spill Conference. Amer. Petro. Inst., Wash., D.C.

A Potter, T.L. 1989. Analysis of Petroleum Contaminated Soil and Water: An Overview. IN: Petroleum Contaminated Soils, Vol. 2, Lewis Publications, Inc. Chelsea, MI

A, Roberts, A.J. and T.C. Thomas. 1986. Characterization and Evaluation of JP-4, Jet A and Mixtures of these Fuels in Environmental Water Samples. Environ. Toxicology and Chemistry. 5:3–11.

A Rygle, K.J. 1987. Methods for "Free" Product Analysis. IN: Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. HESD Dept. Rpt. #214. Amer. Petro. Inst., Wash. D.C.

A, Smith, J.H. et al. 1981. Analysis and Environmental Fate of Air Force Distillate and High Density Fuels. Engineering & Services Lab, Tyndall A.F.B. Available from NTIS, Springfield, VA 703-487-4650. NTIS #AD A115949/LP (Composition and solubility of various military jet fuels)

A Sutton, D.L. 1987. Component Analyses by High Resolution Gas Chromatography. IN: Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. HESD Dept. Rpt. #214. Amer. Petro. Inst., Wash. D.C.

A, Vandegrift, S.A. and D.H. Kampbell. 1988. Gas Chromatographic Determination of Aviation Gasoline and JP-4 Jet Fuel in Subsurface Core Samples. J. Chromatographic Science. 26:566–569.

A Youngless, T.L. et al. 1985. Mass Spectral Characterization of Petroleum Dyes, Tracers and Additives. Analytical Chemistry. 57:1894–1902.

C American Petroleum Institute. 1992. Mineral Oil Review. Health and Envir. Sciences Departmental Report #DR-21. API, 1220 L St. NW, Wash. D.C. 20005

C Bea, D.A. JP-8: Wy, What and When. 1988. Nat. Petro. Refiners Assoc. Fuels & Lubricants Conf. NPRA #FL-88-118.

C Brinkman, D.W. and J. R. Dickson. 1995. Contaminants in Used Lubricating Oils and Their Fate during Distillation / Hydrotreatment Re-Refining. Env. Sci. Tech. 29:81-86

C Canadian Petroleum Products Institute. 1994. Composition of Canadian Summer and Winter Gasolines. CPPI Report No. 94-5. CPPI, Ottawa, ON (Perhaps the definitive gasoline composition collection. Characterizes 128 different gasolines for >40 components of gasoline found in concentrations >1%. Includes summary statistics). CPPI Phone #613/232-3709

C, Chen, C. S., J.J. Delfino & P.S.C. Rao. Partitioning of Organic and Inorganic Components From Motor Oil Into Water. Chemosphere, 28:(7)1385-1400

C Chevron Research and Technology Company. 1990. Motor Gasolines. Lubricant Services, Chevron Res. and Tech. Co., Richmond, CA

C CONCAWE. 1992. Gasolines. Product Dossier no. 92/103. CONCAWE, Brussels, Belgium. (Summarizes health, safety, and environmental data currently available on unformulated gasoline)

C CONCAWE. 1994. Kerosenes / Jet Fuels. Product Dossier no. 94/106. CONCAWE, Brussels, Belgium. (Summarizes health, safety, and environmental data currently available on kerosenes & jet fuels)

C Dickson et al. 1987. Trends in Petroleum Fuels. Nat'l. Inst. for Petroleum and Energy Research, Publ. NIPER-309. Bartlesville, OK (Also, NIPER publishes semi-annual summaries surveys of different fuels that contain data on general fuel characteristics, e.g. volatility, benzene and ether content of gasoline, etc.)

C Domask, W.G. 1984. Introduction to Petroleum Hydrocarbons: Chemistry and Composition in Relation to Petroleum-Derived Fuels and Solvents. IN: Renal Effects of Petroleum Hydrocarbons. Princeton Scientific Publ., Inc. Princeton, N.J.

C, Dunlap, L.E. et al. 1988. Soluble Hydrocarbons Analysis from Kerosene/Diesel Type Hydrocarbons. Vol. I. Proceedings of Petro. Hydrocarbons and Org. Chem. in Groundwater. P. 37–45. Natl. Water Well Assoc., Dublin, OH. (BTEX and naphthalene content of furnace oil, kerosene, diesel, Jet-A; water soluble fraction data also)

C Eccleson, B.H., and F. W. Cox. 1977. Physical Properties of Gasoline/Methanol Mixtures. BERC/RI-76/12. Bartlesville Energy Research Ctr., U.S. Energy Research and Devel. Admin. Tech. Inform. Ctr., Bartlesville, OK. (Appendix A has detailed compositional data for 8 gasolines, all from a single supplier)

C Federal Register. 1987. Crude Oil Analysis, Volume 52, Tuesday, June 16, 1987, p. 22960

C Furey, R.L. 1986. Composition of Vapor Emitted from a Vehicle Gasoline Tank during Refueling. Society of Automotive Engineers Paper 860086

C Gerry, F.S., et al. 1992. Test Fuels: Formulation and Analysis - The Auto/Oil Air Quality Improvement Research Program. Soc. of Automotive Engineers Paper # 920324. Soc. Automotive Eng., Warrendale, PA (Detailed composition of 17 fuels)

C Goodman, D.R., and R.D. Harbison. 1984? Toxicity of the Major Constituents and Additives of Gasoline, Kerosene and No. 2 Fuel Oil. Div. of Interdisciplinary Toxicology, Univ. of Arkansas for Medical Sciences, Little Rock

C Halder, C.A. et al. 1986. Gasoline Vapor Exposures. Part I. Characterization of Workplace Exposures. Amer. Industrial Hygiene Assoc. J. 47(3):164–172. (Data on average composition of gasoline vapors at marketing facilities)

C Johnson, P.C. et al. 1988. Practical Screening Models for Soil Venting Systems. In: Proceedings of Petroleum Hydrocarbons and Organic Chemicals in Groundwater. National Water Well Association, Dublin, OH (p. 544–546; components of Gasoline and weathered gasoline)

C Jokuty, P., S. Whiticar, Z. Wang, M. Fingas, P. Lambert, B. Fieldhouse, and J. Mullin. 1996. A Catalogue of Crude Oil and Oil Product Properties. Manuscript Report EE-157. Environment Canada, Ottawa, Ontario. 956 pages. (Compendium of physical and chemical properties of crude oils, and a limited number of refined products) On the Web at www.ETCentre.org/spills

C, Jordan, R.E. and J.R. Payne. 1980. Fate and Weathering of Petroleum Spills in the Marine Environment. Ann Arbor Science, Ann Arbor, MI. (Comparison of #2 and #6 bunker fuel oil, p.122)

C King, R.W. 1988. Petroleum: Its Composition, Analysis, and Processing. IN: Occupational Medicine: The Petroleum Industry. N.K. Weaver, Ed. Hanley and Belfus, Philadelphia, PA.

C Kreamer, D.K. and Stetzenbach, K.J. 1990. Development of a Standard, Pure-Compound Base Gasoline Mixture for Use as a Reference in Field and Laboratory Experiments. GWMR. Spring 1990. 135–145

C Mayfield, H.T. 1996. JP-8 Composition and Variability. AL/EQ-TR-1996-0006 Armstrong Lab, Tyndall AFB, FL

C Maynard, J.B. and W.N. Sanders. 1969. Determination of the Detailed Hydrocarbon Composition and Potential Atmospheric Reactivity of Full-Range Motor Gasolines. J. Air Poll. Control Assoc. Vol. 19, #7.

C National Research Council. 1985. Oil in the Sea: Inputs, Fate, and Effects. National Academy of Sciences, Washington, D.C. (Composition of crudes, some products, p.19–23)

C Pritchard et al. 1988. Environmental Fate and Effects of Shale-Derived Jet Fuel. Envir. Res. Lab, Gulf Breeze, FL. NTIS #AD-A 197 683/6/WEP. 99 pp.

C Rastogi, S.C. 1993. Residues of Benzene in Chemical Products. Bull. Environ. Contam. Toxicol., 50:794-797.

C, Romeu, A. et al. 1988. Mobilization of Volatile Toxic Components from Petroleum Product-Contaminated Soils by TCLP. IN: Proceedings, U.S. EPA Symposium on Waste Testing and Quality Assurance. July 11–15, 1988, Washington, D.C. Amer. Public Works Assoc., Editors. p. F-25 to F-43. (A condensed paper of the report by Romeu et al. for OUST)

C, Romeu, A., et al. 1988. Determining if Soils Contaminated with Petroleum Products are Hazardous Wastes. Draft Report prepared for the EPA Office of Underground Storage Tanks. Contract # 68-01-7383. April, 1988. UST Docket # UST2-4-SB-29 (Contains data on composition of gasoline, diesel, and #6 fuel oil, plus TCLP leachate values for contaminated soils).

C Shull, L.R. et al. 1994. Development of a Health Risk Assessment Methodology for Mineral Spirits. Hydrocarbon Contaminated Soils: Vol 4 p. 255-274.

C	Sigsby, J.E. et al. 1987. Volatile Organic Compound Emissions from 46 In-Use Passenger Cars. Environ. Sci. Technol. 21:466–475. (composition of no-lead regular and premium gasoline)
C,	Smith, J.H. et al. 1981. Analysis and Environmental Fate of Air Force Distillate and High Density Fuels. Engineering & Services Lab, Tyndall A.F.B. Available from NTIS, Springfield, VA 703-487-4650. NTIS #AD A115949/LP (Composition & solubility of various military jet fuels).
C	Westerholm, R. and H. Li. 1994. A Multivariate Statistical Analysis of Fuel-Related PAH Emissions from Heavy-Duty Diesel Vehicles. Envir. Sci. & Tech. 28: 965-972 (PAH content of diesel)
C	Western States Petroleum Assoc. 1993. Chemical and Physical Characteristics of Crude Oil, Gasoline, and Diesel Fuel: A Comparative Study. WSPA Report. 505 N. Brand Blve, Suite 1400, Glendale, CA 91203
I	Baugh, A., and J. Lovegreen. 1990. Differentiation of Crude Oil and Refined Petroleum Products in Soil. Petroleum Contaminated Soils: Volume 3. Lewis Publishers, Chelsea, MI
I	Bragg, J.R. et al. 1994. Effectiveness of bioremediation for the Exxon Valdez oil spill. Nature, 368: 413-418. (Disc. of use of hopane as a biomarker; crude oil chemistry changes over time)
I	Bruce, L.G. 1993. Refined Gasoline in the Subsurface. Amer. Assoc. of Petrol. Geologists Bulletin, V. 77 #1 p. 142-149.
I	Bruce, L.G. and G. Schmidt. 1994. Hydrocarbon Fingerprinting for Application in Forensic Geology: Review with Case Studies. Amer. Assoc. of Petrol. Geologists Bulletin, V. 78 #11 p. 1692-1710
I, A	Butt, J.A. et al. 1986. Characterization of Spilled Oil Samples: Purpose, Sampling, Analysis and Interpretation. Institute of Petroleum: London, England. (Focus is on crude oil spills in marine environments, but contains good methodology)
I	Christiansen, L.B. and T. Larsen. 1993. Method for Determining the Age of Diesel Spills in the Soil. Ground Water Monitoring & Remediation. Vol. 13, #4, p.142-149.
I, A	Coleman, W.E. et al. 1984. The Identification and Measurement of Components in Gasoline, Kerosine, and No. 2. Fuel Oil that Partition into the Aqueous Phase After Mixing. Arch. Environ. Contam. Toxicol. 13:171–178. (Chromatograms of fuels)
I	Dell'Acqua, R. et al. 1976. Identification of Gasoline Contamination of Groundwater by Gas Chromatography. J. of Chromatography. 128:271–280.
I	Douglas, G.S. et al. 1996. Environmental Stability of Selected Petroleum Hydrocarbon Source and Weathering Ratios. Env. Sci. & Tech. 30:2232-39.
I	Grosser, P.W. and F.P. Castellano. 1993. A Case Study of Petroleum Analysis. Proceedings, 1993 Petroleum Hydrocarbons and Organic Chemicals in Ground Water. p. 189-202. National Ground Water Assoc. Dublin, OH.
I	Hirz, R. 1989. Gasoline Brand Identification and Individualization of Gasoline Lots. J. Forensic Sci. Soc. 29(2): 91-101

I Hughes, B.M., et al. 1989. Examples of the Use of an Advanced Mass Spectrometric Data Processing Environment for the Determination of Sources of Wastes. IN: Proceedings, Waste Testing and Quality Assurance

I Hurst, R.W., T.E. Davis, and B. D. Chinn. 1996. The lead fingerprints of gasoline contamination. Env. Sci. & Tech. 30:304A-307A.

I Ilias, A. M. 1994. Fuel Isolation, Identification and Quantification from Soils. In Proceedings, ASTM Symposium "Analysis of Soils Contaminated with Petroleum Constituents". p. 12-26. O'Shay & Hoddinott, Eds. ASTM Publ. STP 1221. ASTM, Phil. PA

I Kanai, J., V. Inouye, R. Goo, L. Yasawa, J. Maka, and C. Chun. 1991. Gas Chromatographic/Mass Spectrometric Analysis of Polar Components in "Weathered" Gasoline/Water Matrix as an Aid in Identifying Gasoline. Anal. Letters. 24(1):115-128

I Kaplan, I.R. 1996. Fingerprinting and Age Dating of Hydrocarbon Releases: High Boiling Fuels, Asphalts and Lubricants. Hazmat West '96 Conference Proceedings.

I Kaplan, I.R. and Y. Galperin. 1996. Determining the Age of Hydrocarbon Fuel Spills Using Organic Geochemical Data.

I Kaplan, I.R. and Y. Galperin. 1996. How to Recognize a Hydrocarbon Fuel in the Environment and Estimate Its Age of Release. Chapter 8 in: "Groundwater and Soil Contamination: Technical Preparation and Litigation Management", T. Bois and B. Luther, Eds. John Wiley & Sons

I Kaplan, I.R. 1992. Characterizing Petroleum Contaminants in Soil and Water and Determining Source of Pollutants. In Proceedings of: 1992 Petroleum Hydrocarbons in Groundwater Conference. p. 3-18. API/NGWA. Nat. Ground Water Assoc. Dublin, OH.

I Kaplan, I.R. 1992. Environmental Forensic Geochemistry: A Chemical System for Identifying Sources of Escaped Petroleum Products. In Proceedings of: HAZMAT West '92. Long Beach, CA

I Kaplan, I.R., Y. Galperin, H. Alimi, R. Lee and S. Lu. 1996. Patterns of Chemical Changes During Environmental Alteration of Hydrocarbon Fuels. Ground Water Mon. & Remed. 16:4, 113-124

I Kirkbride, K.P., S.M. Yap et al. 1992. Microbial Degradation of Petroleum Hydrocarbons: Implication for Arson Residue Analysis. J. of Forensic Science. 6:1585-1599.

I Kitto, M.E. 1993. Emission of Rare-Earth Elements (REE) from Oil-Related Industries. Air & Waste Mngmt. Assoc. Annual Mtg, Paper #93TP57.01, 13 p.

I Kvenvolden, K.A. and P.R. Carlson. 1994. Carbon Isotopic Identification of Two Sources of Oil Residues in Prince William Sound, AK. Amer. Chem. Soc. Mtg. Abstract, Mar. 1994

I, Lavine, B.K., H. Mayfield, P.R. Kromann and A. Faruque. 1995. Source Identification of Underground Fuel Spills by Pattern Recognition Analysis of High-Speed Gas Chromatograms. Analy. Chem. 67:3846-3852.

Lesage, S., H.Xu, and K.S. Novakowski. 1997. Distinguishing Natural Hydrocarbons from Anthropogenic Contamination in Ground Water. Ground Water 35 (1):149-160.

Luhrs, R.C. and C.J. Pyott. 1992. Trilinear Plots: A Powerful New Application for Mapping Gasoline Contamination. In Proceedings of: 1992 Petroleum Hydrocarbons in Groundwater Conference. p. 85-102 API/NGWA. Nat. Ground Water Assoc. Dublin, OH.

Mann, D.C., and W.R. Gresham. 1990. Microbial Degradation of Gasoline in Soil. J. of Forensic Sciences. 4:913-923.

Mansuy, L., R.P. Philp and J. Allen. 1997. Source identification of oil spills based on the isotopic composition of individual components in weathered oil samples. Env. Sci. & Tech., 31:3417-25

Petroleum Assoc. for Conservation of the Canadian Environment. 1990. Weathering of Crude Oil in Surface Soils. PACE Rpt. # 90-3. CPPI, Ottawa, Ontario, Canada

Pharr, D.Y., Mckenzie, J.K., and A.B. Hickman. 1992. Fingerprinting Petroleum Contamination Using Synchronous Scanning Fluorescence Spectroscopy. Groundwater, 30:484-489.

Pollard, S.J. T. et al. 1994. A Tiered Analytical Protocol for the Characterization of Heavy Oil Residues at Petroleum Contaminated Hazardous Waste Sites. In Proceedings, ASTM Symposium "Analysis of Soils Contaminated with Petroleum Constituents". p. 38-52. O'Shay & Hoddinott, Eds. ASTM Publ. STP 1221. ASTM, Phil. PA

Potter, T. 1990. Fingerprinting Petroleum Products: Unleaded Gasoline. Petroleum Contaminated Soils: Volume 3. Lewis Publishers, Chelsea, MI

Reyes, M.V. 1991. Derivative Spectroscopy as an Analytical Tool for Hydrocarbon Identification. Symp. on Modern Analytical Tech. for the Analysis of Petroleum. Amer. Chem. Soc., New York City mtg., August, 1991. Preprints, Vol. 36, #2: p. 291-304.

Roques, D.E. E.B. Overton and C.B. Henry. 1994. Using Gas Chromatograph/Mass Spectroscopy Fingerpring Analysis to Document Process and Progress of Oil Degradation. J. of Envir. Qual. 23:851-855

Saenz, G., And N.E. Pingitore. 1991. Characterization of Hydrocarbon Contaminated Areas by Multivariate Statistical Analysis Case Studies. Envir. Mon. and Assessment 17:281-302

Senn, R.B. and M.S. Johnson. 1987. Interpretation of Gas Chromatographic Data in subsurface Hydrocarbon Investigations. Groundwater Monitoring Review, Winter, 1987; p. 58–63.

Siddiqui, K.J., R. L. Lidberg, D. Eastwood, and G. Gibson. 1991. Expert Systems for Classification and Identification of Waterborne Oils. IN: "Monitoring Water in the 1990's: Meeting New Challenges." ASTM STP 1102. Hall and Glysson, Eds. ASTM, Philadelphia, PA

Steiber, R.S. 1993. Organic Combustion Fingerprints of Three Common Home Heating Fuels. J. Air & Waste Mngmt. v. 43 p. 859-863

I	Testa, S.M. and W.E. Halbert. 1989. Geochemical Fingerprinting of Free Phase Liquid Hydrocarbon. IN: Proceedings, Petro. Hydrocarbons and Org. Chem. in Groundwater. Natl. Water Well Assoc., Dublin, OH.
I	Thomas, D.H., and J. J. Delfino. 1991 A Gas Chromatographic/Chemical Indicator Approach to Assessing Ground Water Contamination by Petroleum Products. Groundwater Monitoring Review, Fall, 1991: p. 90-100
I	Wang, Z. M. and M. Fingas. 1995. Use of Methyldibenzothiophenes as Markers for Differentiation and Source Identification of Crude and Weathered Oils. Env. Sci. Tech. 29: 2842-49.
I	Wang, Z. M. Fingas and G. Sergy. 1994. Study of 22-year Old Arrow Oil Samples Using Biomarker Compounds by GCMS. Env. Sci. Tech. 28: 1733-46
I	Whittemore, I.M. 1987. Identification of Spilled Hydrocarbons. **IN:** Proceedings, Sampling and Analytical Methods for Determining Petroleum Hydrocarbons in Groundwater and Soil. HESD Dept. Rpt. #214. Amer. Petro. Inst., Wash. D.C.
I	Worthington, M.A. and E.J. Perez. 1993. Dating Gasoline Releases Using Ground Water Chemical Analyses. Proceedings, 1993 Petroleum Hydrocarbons and Organic Chemicals in Ground Water. 203-220. National Ground Water Assoc. Dublin, OH.
I	Zemo, D.A., T.E. Graf, and J.E. Bruya. 1993. The Importance and Benefit of Fingerprint Characterization in Site Characterization and Remediation, Focusing on Petroleum Hydrocarbons. Proceedings, 1993 Petroleum Hydrocarbons and Organic Chemicals in Ground Water. 55-72. National Ground Water Assoc. Dublin, OH.
I, A	Zerlia, T., et al. 1990. UV Spectrometry as a Tool for Rapid Screening of Petroleum Products. FUEL; 69:1381-85.
S,	API 1985. Literature Survey: Hydrocarbon Solubilities and Attenuation Mechanisms, Publication #4414.
S,	API. 1985. Laboratory Study on Solubilities of Petroleum Hydrocarbons. API Publ. #4395. ., Wash. D.C. G.T. Brookman et al., authors
S,	API. 1991. Solubility of BTEX from Gasoline/Oxygenate Mixtures. Publ. 4531. Amer. Petroleum Institute, Wash. D.C. M. Poulson et al., authors
S	Burchette, G. 1986. No. 6 Fuel Oil and Ground Water. Ground Water Monitoring Review. 6:32. (A brief letter noting the presence of BTEX in #6 fuel oil.)
S, I	Chen, C. S-H., J.J. Delfino, and P.S. Rao. 1994. Partitioning of Organic and Inorganic Components from Motor Oil into Water. Chemosphere, 28:(7) 1385-1400.
S	Cline, P.V., J.J. Delfino, and P.S. Rao. 1991. Partioning of Aromatic Constituents into Water from Gasoline and Other Complex Solvent Mixtures. Environ. Sci. Technol.; 26:5, 914-920.
S,	Coleman, W.E. et al. 1984. The Identification and Measurement of Components in Gasoline, Kerosene, and No. 2. Fuel Oil that Partition into the Aqueous Phase After Mixing. Arch. Environ. Contam. Toxicol. 13:171–178. (Chromatograms of fuels)

S	Guard, H.E. et al. 1983. Characterization of Gasolines, Diesel Fuels, and Their Water Soluble Fractions. Naval Biosciences Laboratory, Naval Supply Center, Oakland, CA. Avail from NTIS: AD-A270 016 (Water sol. fraction of leaded, unleaded, and premium gasolines, and diesel)
S	Kramer, W.H. and T.J. Hayes. 1987. Water Soluble Phase of Gasoline: Results of a Laboratory Mixing Experiment. New Jersey Geological Survey Tech. Mem. 87-5. NJGS, Trenton, NJ
S	Kramer, W.H. and T.J. Hayes. 1987. Water Soluble Phase of Number 2 Fuel Oil: Results of a Laboratory Mixing Experiment. New Jersey Geol. Survey Tech. Mem. 87-4. NJGS, Trenton, NJ
S	Lane, W.F. and R.C. Loehr. 1992. Estimating the Equilibrium Aqueous Concentrations of Polynuclear Aromatic Hydrocarbons in Complex Mixtures. Environ. Sci. Technol. 26:983-990.
S	Lee, L.S., M.Hagwell, J. J. Delfino and P.S.C. Rao. 1992. Partioning of PAHs from diesel fuel into water. Envir. Sci. & Tech. 26:2104-2110.
S	Lu, B.C.Y. and Jiri Polak. 1976. A Study of the Solubility of Oil in Water. Technology Development Report EPS-4-EC-76-1. Environment Canada, Ottawa, Canada (Data on No. 2 fuel oil, crude oil, and medium bunker fuel)
S	Mackay, D., W.Y. Shiu, A. Maijanen, and S. Feenstra. 1991. Dissolution of non-aqueous phase liquids in groundwater. J. of Contaminant Hydrology, 8:23-42.
S,	Potter, T. 1992. Aqueous Solubility and Analysis of Gasoline. Ph.D disseration, U. Massachusetts. UMI Dissertation Information Service, Ann Arbor, MI.
S	Shiu, W.Y., M. Bobra, A. Bobra, A. Maijanen, L. Suntio, and D. Mackay. 1990. The Water Solubility of Crude Oils and Petroleum Products. Oil & Chemical Pollution, 7:57-84.

Further References from American Petroleum Institute

ADDITIVES

Cummings, W.M. 1977. Fuel and Lubricant Additives. Lubrication, 63:1–12

Gibbs, L.M. 1989. Additives Boost Gasoline Quality. Oil & Gas Journal. Apr. 24, 60–63.

Gibbs, L.M. 1990. Gasoline Additives - When and Why. SAE Tech. Paper #902104. Int'l. Fuels and Lubricants Meeting, Oct. 1990. Soc. Automotive Eng., Warrendale, PA

Mansell, R.S., L. Ou, R.D. Rhue and Y. Ouyang. 1995. The fate and behavior of lead alkyls in the subsurface environment. AL/EQ-TR-1994-0026. Armstrong Lab, Tyndall AFB, FL.

NTIS. 1989. Diesel Fuel Additives. Petroleum Review. October, 35–42.

Pipinger, G. 1996. Making "Premium" Diesel Fuel. Hydrocarbon Processing, Feb. 1997 63-77. (Review of different diesel additives)

Ranney, M.W. Fuel Additives for Internal Combustion Engines.

Rhue, R.D., R.S. Mansell, L.T. Ou, R. Cox, S.R. Tang, and Y. Ouyang. 1992. The Fate and Behavior of Lead Alkyls in the Environment: A Review. Crit. Reviews in Env. Control, 22:169-193.

Russell, T.J. 1988. Petrol and Diesel Additives. Petroleum Review. October, 35–42.

Tupa, R.C. and C.J. Dorer. 1984. Gasoline Additives for Performance/Distribution/Quality. Society for Automotive Engineers Paper 841211. Soc. Automotive Eng., Warrendale, PA

GENERAL REFERENCES ON FUELS

Altgelt, K.H. and M.M. Boduszynski. 1996. Composition and Analysis of Heavy Petroleum Fractions. 512 pp. Marcel Dekker, NY, NY (212/696-9000)

ASTM. 1993. Manual on the Significance of Tests for Petroleum Products. 6th Edition. ASTM Manual Series MNL 1. G.V. Dyroff, Ed. ASTM, Philadelphia, PA [This deserves special mention for its completeness, addressing the total range of products, with very good summaries of their general composition and characteristics.]

ASTM. 1993. Manual on Hydrocarbon Analysis. 5th Edition. ASTM Manual Series MNL 3. A.W. Drews, Ed. ASTM, Philadelphia, PA

ASTM. 1993. ASTM and Other Specifications and Classifications for Petroleum Products and Lubricants. 6th Edition. ASTM, Philadelphia, PA

Camin, D.L. 1979. History of Chromatography in Petroleum Analysis. In "Chromatography in Petroleum Analysis", Altgelt and Gouw, eds. Marcel Dekker, NY, NY p. 1-12

CONCAWE. 1992. Gasolines. Product Dossier #92/103. Petroleum Products and Health Management Groups, CONCAWE, Brussels, Belgium

CONCAWE. 1995. Kerosenes / Jet Fuels. Product Dossier #94/106. Petroleum Products and Health Management Groups, CONCAWE, Brussels, Belgium

CONCAWE. 1995. Gas Oils (diesel fuels / heating oils). Product Dossier #95/107. Petroleum Products and Health Management Groups, CONCAWE, Brussels, Belgium

Garret, T.K. 1991. Automotive Fuels and Fuel Systems. Soc. of Automotive Eng., Warrendale, PA

Guthrie, V. B. (editor). 1960. Petroleum Products Handbook. McGraw-Hill Publ., New York (Out of Print, but has a lot of very good information)

Orszulik, S.T. and R. M. Mortier. 1992. Chemistry and Technology of Lubricants. VCH Publishers, Deerfield Beach, FL.

Owen, K., and T. Coley. 1990. Automotive Fuels Handbook. Society of Automotive Engineers. Warrendale, PA

Speight, J.K. 1991. Chemistry and Technology of Petroleum, 2nd Edition. Marcel Dekker, Inc. New York, NY

Results from BIDS searches on oil, chromatography, neural network, pattern recognition

Only the most recent (past 5 years) papers are included.
Only English language papers are included.
Presented in standard format from Bath Information and Data Services (BIDS); results from ISI, EI and RSC databases.

1. Comparisons of neural network and statistical pattern recognition techniques

TI: Chromatography pattern recognition of Aroclors using iterative probabilistic neural networks.
AU: Magelssen_GR, Elling_JW
JN: Journal of Chromatography, A, 1997, Vol.775, No.1-2, pp.231-242
IS: 0021-9673
DT: Article
NA: Los Alamos Natl. Lab., Los Alamos, NM 87545, USA
CO: Presented at the 1st SFE/SFC/XSE Symposium, held in Siegen, Germany, 1-2 Oct 1997
AB: Standard Aroclor mixtures, e.g., Aroclors 1242, 1254 and 1260 (0.05-0.8 mug/ml) were chromatographed on a column (20 m * 0.25 mm i.d.) coated with DB-1 (0.1 mum), with a carrier gas flow rate of 1.8 ml/min) and temperature programming from 80-320degC (no other details given). Mixtures of the three Aroclors were also analysed and random noise was generated as chromatograms. The results were analysed by standard data processing methods, by linear regression pattern recognition and by iterative probabilistic neural network as methods of classifying the results on environmental samples. The neural network method, which provides percentage probabilities of identification, gave more accurate and more sensitive results than the other two in identifying single compounds in mixtures of Aroclor-free samples. In particular, it was free from the false positives given by the other methods from random noise.

TI: Clustering of infrared spectra of lubricating base oils using adaptive resonance theory.
AU: Wang_XZ, Chen_BH
JN: Journal of Chemical Information and Computer Science, 1998, Vol.38, No.3, pp.457-462
IS: 0095-2338
DT: Article
NA: Dept. Chem. Eng., Univ. Leeds, Leeds LS2 9JT, UK
AB: An unsupervised neural network, the adaptive resonance theory, was applied to the classification of IR spectra for 59 data samples from 12 refineries representing 8 crude oil origins. This approach does not require known and classified data sets. Seven classes were identified, five of which perfectly matched the crude oil origins. The results were similar to those obtained with principal-components analysis.

TI: Soy sauce classification by geographic region based on NIR spectra and chemometrics pattern recognition
AU: Iizuka_K, Aishima_T
NA: KIKKOMAN FOODS INC, 399 NODA, NODA, CHIBA 278, JAPAN
JN: JOURNAL OF FOOD SCIENCE, 1997, Vol.62, No.1, p.101 (5 pages)
IS: 0022-1147
DT: Article
AB: Statistical and artificial neural network (ANN) pattern recognition techniques were applied to NIR spectra of 38 soy sauce samples collected from the northern/central, western, and southern regions in Japan and related to differences in food flavorings. Linear discriminant analysis (LDA) and ANN using factor scores calculated from NIR spectra showed more accurate differentiations than those based on the original spectra. In LDA, the correctly assigned ratio was 81.6%. Correct classification ratios shown by Partial least squares (PLS2) were 84.2% and by ANN 76.3% in the cross-validation test. The differentiations suggested that there are quality differences in soy sauce among the three regions in Japan.

TI: Rapid identification of Streptococcus and Enterococcus species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks
AU: Goodacre_R, Timmins_EM, Rooney_PJ, Rowland_JJ, Kell_DB
NA: UNIV WALES, INST BIOL SCI, ABERYSTWYTH SY23 3DA, DYFED, WALES; YSBYTY CYFFREDINOL BRONGLAIS BRONGLAIS GEN HOSP, ABERYSTWYTH SY23 1ER, DYFED, WALES; UNIV WALES, DEPT COMP SCI, ABERYSTWYTH SY23 3DB, DYFED, WALES
JN: FEMS MICROBIOLOGY LETTERS, 1996, Vol.140, No.2-3, pp.233-239
IS: 0378-1097
DT: Article
AB: Diffuse reflectance-absorbance Fourier transform infrared spectroscopy (FT-IR) was used to analyse 19 hospital isolates which had been identified by conventional means to one of Enterococcus faecalis, E. faecium, Streptococcus bovis, S. mitis, S. pneumoniae, or S. pyogenes. Principal components analysis of the FT-IR spectra showed that this 'unsupervised' learning method failed to form six separable clusters (one for each species) and thus could not be used to identify these bacteria based on their FT-IR spectra. By contrast, artificial neural networks (ANNs) could be trained by 'supervised' learning (using the back-propagation algorithm) with the principal components scores of derivatised spectra to recognise the strains from their FT-IR spectra. These results demonstrate that the combination of FT-IR and ANNs provides a rapid, novel and accurate bacterial identification technique.

TI: Classification of microbial defects in milk using a dynamic headspace gas chromatograph and computer-aided data processing
AU: Horimoto_Y, Lee_K, Nakai_S
NA: UNIV BRITISH COLUMBIA, DEPT FOOD SCI, 6650 NW MARINE DR, VANCOUVER, BC V6T 1Z4, CANADA
JN: JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY, 1997, Vol.45, No.3, pp.743-747
IS: 0021-8561
DT: Article
AB: Objective, yet cost-effective evaluation of flavor is difficult in quality control of milk. Inexpensive gas chromatographs in conjunction with computer models make it feasible to construct an objective flavor evaluation system far routine quality control purposes. The purpose of this study was to classify milk with microbial off-flavors using a low-cost headspace gas chromatograph and computer-aided data processing. Principal component similarity (PCS) analysis was discussed in part 1. In part 2, artificial neural networks (ANN), partial least-squares regression (PLS) analysis, and principal component regression (PCR) analysis are examined. UHT milk was inoculated with various bacteria (Pseudomonas fragi, Pseudomonas fluorescens, Lactococcus lactis, Enterobactor aerogenes, and Bacillus subtilis) and a mixed culture (P. fragi:E. aerogenes:L. lactis = 1:1:1) to approximately 4.0 log(10) CFU mL(-1). ANN were able to make better predictions than PLS and PCR. The prediction ability of PLS was better than PCR. The performance of each method depended on the content of training and testing of data, i.e., more data resulted in better predictive ability.

TI: Classification of wine samples by means of artificial neural networks and discrimination analytical methods
AU: Sun_LX, Danzer_K, Thiel_G
NA: UNIV JENA, INST INORGAN & ANALYT CHEM, D-07743 JENA, GERMANY
JN: FRESENIUS JOURNAL OF ANALYTICAL CHEMISTRY, 1997, Vol.359, No.2, pp.143-149
IS: 0937-0633
DT: Article
AB: The three-layer artificial neural network (ANN) model with back-propagation (BP) of error was used to classify wine samples in six different regions based on the measurements of trace amounts of B, V, Mn, Zn, Fe, Al, Cu, Sr, Ba, Rb, Na, P, Ca, Mg, K using an inductively coupled plasma optical emission spectrometer (ICP-OES). The ANN architecture and parameters were optimized. The results obtained with ANN were compared with those obtained by cluster analysis, principal component analysis, the Bayes discrimination method and the Fisher discrimination method. A satisfactory prediction result (100%) by an artificial neural network using the jackknife leave-one-out procedure was obtained for the classification of wine samples containing six categories.

TI: Plant seed classification using pyrolysis mass spectrometry with unsupervised learning: The application of auto-associative and Kohonen artificial neural networks
AU: Goodacre_R, Pygall_J, Kell_DB
NA: UNIV WALES, INST BIOL SCI, ABERYSTWYTH SY23 3DA, DYFED, WALES
JN: CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 1996, Vol.34, No:1, pp.69-83
IS: 0169-7439
DT: Article
AB: Pyrolysis mass spectrometry (PyMS) was used to gain high dimensional (150 m/z values) biochemical fingerprints from Begonia semperflorens Summer Rainbow, Campanula carpatica White Gem, Lobelia erinus White Fountain, and Lobelia erinus White Lady plant seeds, Rather than homogenizing the seeds and analysing the extracts, the sample preparation of the seeds in this study was novel and merely involved crimping the metal foil sample carrier around the seeds. Compared to extractive procedures the technique exploited in this study will give a fair representation of the seed, is rapid and thus amenable to the analysis of a high volume of samples. To observe the relationship between these seeds, based on their spectral fingerprints, it was necessary to reduce the dimensionality of these data by unsupervised feature extraction methods. The neural computational pattern recognition techniques of self organising feature maps (SOFMs) and auto-associative neural networks were therefore employed and the clusters observed compared with the groups obtained from the more conventional statistical approaches of principal components analysis (PCA) and canonical variates analysis (CVA). When PCA was used to analyze the raw pyrolysis mass spectra replicate samples were not recovered in discrete clusters; CVA, which minimises the within-group variance and maximises the between-group variance, therefore had to be employed. Although B. semperflorens and C. carpatica seeds were recovered separately and away from the L. erinus plant seeds, the two types of L. erinus seeds could still not be discriminated between using this approach. CVA uses a priori information on which spectra are replicates; we therefore encoded this information by employing a novel preprocessing regime where the triplicate mass spectra from each of the seeds were averaged in pairs to produce three new spectra; these were then used by each of the unsupervised methods. PCA still failed to separate the two L. erinus; however, auto-associative neural networks could be used successfully to discriminate them. It is likely that this was due to their ability to perform non-linear mappings and hence approximate non-linear PCA. SOFMs could also be used to separate all four seeds unequivocally. To obtain quantitative information regarding the similarity of these seeds from their pyrolysis mass spectra, SOFMs were trained with different numbers of nodes in the Kohonen output layers. The results observed from this procedure are often difficult to report in tables or visualise using topological contour maps; to simplify the graphical representation of the similarity between the seeds we therefore performed the novel construction of a dendrogram from the various SOFMs analyses. This study demonstrates the potential of PyMS for discriminating plant seeds at the genus, species and sub-species level. Moreover the clusters observed were a true reflection of the known taxonomy of these plants. This approach will be invaluable to the plant taxonomist in representing biological relationships among plant taxa or in describing genomic relationships without the need for cultivation of the propagule.

2. Neural networks and other AI techniques

TI: Knowledge discovery in databases; application to chromatography.
AU: Bryant_CH, Rowe_RC
JN: Trends in Analytical Chemistry, 1998, Vol.17, No.1, pp.18-24
IS: 0165-9936
DT: Article
NA: School Computing and Mathematics, Univ. Huddersfield, Huddersfield HD1 3DH, UK
AB: A review is presented of knowledge discovery in databases (KDD). KDD is based on statistics, pattern recognition, databases, data visualization and areas of artificial intelligence such as machine learning, machine discovery and knowledge acquisition for expert systems. The application of these machine learning techniques to chromatography (liquid, gas or ion) is discussed. (22 references).

TI: Fuel identification by neural network analysis of the response of vapour-sensitive sensor arrays.
AU: McCarrick_CW, Ohmer_DT, Gilliland_LA, Edwards_PA, Mayfield_HT
JN: Analytical Chemistry, 1996, Vol.68, No.23, pp.4264-4269
IS: 0003-2700
DT: Article
NA: Dept. Chem., Edinboro Univ. Pennsylvania, Edinboro, PA 16444, USA
AB: A 2 mul sample of jet fuel was injected into a stream (15 l/min) of air through a GC injection port at sime100degC into a PVC sample chamber (sime20 l) fitted with eight vapour-sensitive sensors and a mixer propeller. The resistance of each sensor, set for appropriate levels (listed) of benzene, toluene, diesel oil, gasoline, JP-4 and JP-5 aviation fuels, ethylbenzene and fuel oil, was monitored with a multichannel control unit. The response of each detector was averaged and stored as the overall response for each sample. Visual inspection of each overall response revealed a characteristic pattern in the response of the array to five of the six different fuel types, and this was confirmed with neural network analyses of the entire data set. Initially fuels were separated into one of five groups, viz, JP-4, JP-5, JP-7, AvGas and a combined JP-8/JetA group. In a second step, fuels in the combined group were separated into either JP-8 or JetA groups.

TI: Sensory evaluation of virgin olive oils by artificial neural network processing of dynamic headspace gas-chromatographic data.
AU: Angerosa_F, DiGiacinto_L, Vito_R, Cumitini_S
JN: Journal of the Science of Food and Agriculture, 1996, Vol.72, No.3, pp.323-328
IS: 0022-5142
DT: Article
NA: Ist. Sperimentale Elaiotecnica, 65013 Citta S Angelo, Pescara, Italy
AB: Olive oil (50 g) was mixed with 7 mg nonan-1-ol (internal standard). Volatiles were stripped with N2 gas at a flow rate of 20 ml/s at 37degC for 2 h then concentrated on to 50 mg activated charcoal (20-30 mesh) and eluted with 1 ml diethyl ether. Analysis was carried out on a 0.5 mum Carbowax 20 m column (50 * 0.32 mm i.d.). Temperature programming was from 25degC (held for 7 min) to 33degC (no hold) at 0.8degC/min, then to 80degC (no hold) at 2.4degC/min and to 155degC (held for 20 min) at 3.7degC/min. H2 was the carrier gas at 30 kPa. FID was used. Results were analysed with an artificial neural network using the back propagation algorithm. Results correlated well with those obtained by sensory evaluation by a panel test (details given). The neural network was able to generalize well and to assign sensory evaluations with a good degree of accuracy.

TI: A fuzzy adaptive resonance theory supervised predictive mapping neural network applied to the classification of multivariate chemical data
AU: Song_XH, Hopke_PK, Bruns_MA, Bossio_DA, Scow_KM
NA: CLARKSON UNIV, DEPT CHEM, POTSDAM, NY, 13699 UNIV CALIF DAVIS, DEPT LAND AIR & WATER RESOURCES, DAVIS, CA, 95616
JN: CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 1998, Vol.41, No.2, pp.161-170
IS: 0169-7439
DT: Article
AB: A fuzzy adaptive resonance theory-supervised predictive mapping (Fuzzy ARTMAP) neural network has been studied for the classification of multivariate chemical data. Fuzzy ARTMAP achieves a synthesis of fuzzy logic and adaptive resonance theory (ART) by exploiting the close formal similarity between the computations of fuzzy subset membership and ART category choice, resonance, and learning. To examine the properties of Fuzzy ARTMAP, the well-known Italian olive oil data set was employed. Then this method was applied to a practical agricultural data set to classify different soil samples depending on the crops grown on them. For comparison, the back-propagation (BP) neural network has also-been used to treat these data. The results show that the classification performance of the Fuzzy ARTMAP neural network is as good or better than the BP network in the present applications. Among other features, the Fuzzy ARTMAP needs less training time and fewer algorithmic parameters to be optimized than BP does to achieve good classification.

TI: NEURAL-NETWORK CLASSIFICATION OF WHEAT USING SINGLE KERNEL NEAR-INFRARED TRANSMITTANCE SPECTRA
AU: SONG_HP, DELWICHE_SR, CHEN_YR
NA: USDA ARS, BELTSVILLE AGR RES CTR, INSTRUMENTAT & SENSING LAB, BELTSVILLE, MD, 20705
JN: OPTICAL ENGINEERING, 1995, Vol.34, No.10, pp.2927-2934
IS: 0091-3286
DT: Article
AB: To investigate an accurate, rapid, and nondestructive method for wheat classification in inspection terminals, backpropagation neural network models were developed, based on single wheat kernel near-infrared transmittance spectra. Six classes of wheat were studied. Neural network models were optimized for two-class and six-class classification. The wavelength range of the spectra was 850 to 1049 nm. For two-class models with 200 input nodes, the average classification accuracy was 97% to 100%. For the six-class model with 200 input nodes, the average accuracy was 94.7%. The classification between hard red winter (HRW) and hard red spring (HRS) was least accurate among the six classes. For rapid classification, a narrower wavelength range, 899 to 1049 nm, with an interval of 2 nm, was proposed and shown to have little loss in accuracy, The most time-consuming two-class (HRW-HRS) model could be calibrated and validated in less than 7 min. Prediction for new data was nearly instantaneous. A backpropagation neural network model with a learning coefficient of 0.6 to 0.65 and momentum of 0.4 to 0.45, without a hidden layer, was effective for wheat classification.

TI: Orange juice classification with a biologically based neural network
AU: Dettmar_HP, Barbour_GS, Blackwell_KT, Vogl_TP, Alkon_DL, Fry_FS, Totah_JE, Chambers_TL
NA: ENVIRONM RES INST MICHIGAN, ARLINGTON, VA, 22209 NINCDS, BETHESDA, MD, 20892; US FDA, CTR FOOD SAFETY & APPL NUTR, WASHINGTON, DC, 20204
JN: COMPUTERS & CHEMISTRY, 1996, Vol.20, No.2, pp.261-266
IS: 0097-8485
DT: Article
AB: Dystal, an artificial neural network, was used to classify orange juice products. Nine varieties of oranges collected from six geographical regions were processed into single-strength, reconstituted or frozen concentrated orange juice. The data set represented 240 authentic and 173 adulterated samples of juices; 16 variables [8 flavone and flavanone glycoside concentrations measured by high-performance liquid chromatography (HPLC) and 8 trace element concentrations measured by inductively coupled plasma spectroscopy] were selected to characterize each juice and were used as input to Dystal. Dystal correctly classified 89.8% of the juices as authentic or adulterated. Classification performance increased monotonically as the percentage of pulpwash in the sample increased. Dystal correctly identified 92.5% of the juices by variety (Valencia vs non-Valencia).

TI: The use of neural networks for fitting complex kinetic data
AU: Galvan_IM, Zaldivar_JM, Hernandez_H, Molga_E
NA: COMMISS EUROPEAN COMMUNITIES, JOINT RES CTR, INST SAFETY TECHNOL, PROC ENGN DIV, TP 680, I-21020 ISPRA, VA, ITALY; WARSAW UNIV TECHNOL, DEPT CHEM & PROC ENGN, PL-00645 WARSAW, POLAND; COMMISS EUROPEAN COMMUNITIES, JOINT RES CTR, INST PROSPECT TECHNOL STUDIES, SEVILLE 41092, SPAIN
JN: COMPUTERS & CHEMICAL ENGINEERING, 1996, Vol.20, No.12, pp.1451-1465
IS: 0098-1354
DT: Article
AB: In this paper the use of neural networks for fitting complex kinetic data is discussed. To assess the validity of the approach two different neural network architectures are compared with the traditional kinetic identification methods for two cases: the homogeneous esterification reaction between propionic anhydride and 2-butanol, catalysed by sulphuric acid, and the heterogeneous liquid-liquid toluene mononitration by mixed acid. A large set of experimental data obtained by adiabatic and heat flux calorimetry and by gas chromatography is used for the training of the neural networks. The results indicate that the neural network approach can be used to deal with the fitting of complex kinetic data to obtain an approximate reaction rate function in a limited amount of time, which can be used for design improvement or optimisation when, owing to small production levels or time constraints, it is not possible to develop a detailed kinetic analysis.

TI: Analysis of accelerants and fire debris using aroma detection technology
AU: Barshick_SA
NA: OAK RIDGE NATL LAB, POB 2008, OAK RIDGE, TN, 37831
JN: JOURNAL OF FORENSIC SCIENCES, 1998, Vol.43, No.2, pp.284-293
IS: 0022-1198
DT: Article
AB: The purpose of this work was to investigate the utility of electronic aroma detection technologies for the detection and identification of ignitable liquid accelerants and their residues in suspected arson debris. Through the analysis of "known" accelerants and residues, a trained neural network was developed for classifying fire debris samples. Three "unknown" items taken from actual fire debris that had contained the fuels, gasoline, kerosene, and diesel fuel, were classified using this neural network. One item, taken from the area known to have contained diesel fuel, was correctly identified as diesel fuel residue every time. For the other two "unknown" items, variations in sample composition, possibly due to the effects of weathering or increased sample humidities, were shown to influence the sensor response. This manifested itself in inconsistent fingerprint patterns and incorrect classifications by the neural network. Sorbent sampling prior to aroma detection was demonstrated to reduce these problems and allowed improved neural network classification of the remaining items which were identified as kerosene and gasoline residues.

TI: Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks
AU: Goodacre_R, Timmins_EM, Burton_R, Kaderbhai_N, Woodward_AM, Kell_DB, Rooney_PJ
NA: UNIV WALES, INST BIOL SCI, ABERYSTWYTH SY23 3DD, WALES; BRONGLAIS GEN HOSP, ABERYSTWYTH SY23 1ER, WALES
JN: MICROBIOLOGY-UK, 1998, Vol.144, No.Pt5, pp.1157-1170
IS: 1350-0872
DT: Article
AB: Three rapid spectroscopic approaches for whole-organism fingerprinting pyrolysis mass spectrometry (PyMS), Fourier transform infra-red spectroscopy (FT-IR) and dispersive Raman microscopy - were used to analyse a group of 59 clinical bacterial isolates associated with urinary tract infection. Direct visual analysis of these spectra was not possible, highlighting the need to use methods to reduce the dimensionality of these hyperspectral data. The unsupervised methods of discriminant function and hierarchical cluster analyses were employed to group these organisms based on their spectral fingerprints, but none produced wholly satisfactory groupings which were characteristic for each of the five bacterial types. In contrast, for PyMS and FT-IR, the artificial neural network (ANN) approaches exploiting multi-layer perceptrons or radial basis functions could be trained with representative spectra of the five bacterial groups so that isolates from clinical bacteriuria in an independent unseen test set could be correctly identified. Comparable ANNs trained with Raman spectra correctly identified some 80% of the same test set. PyMS and FT-IR have often been exploited within microbial systematics, but these are believed to be the first published data showing the ability of dispersive Raman microscopy to discriminate clinically significant intact bacterial species. These results demonstrate that modern analytical spectroscopies of high intrinsic dimensionality can provide rapid accurate microbial characterization techniques, but only when combined with appropriate chemometrics.

3. Statistical pattern recognition techniques.

TI: Gas-chromatographic amino-acid profiling of wine samples for pattern recognition.
AU: Kim_KR, Kim_JH, Cheong_EJ, Jeong_CM
JN: Journal of Chromatography, A, 1996, Vol.722, No.1-2, pp.303-309
IS: 0021-9673
DT: Article
NA: Coll. Pharm., Sungkyunkwan Univ., Suwon 440-746, South Korea
CO: Presented at the International Symposium on Chromatography on the occasion of the 35th anniversary of the Research Group on Liquid Chromatography in Japan, held in Yokohama, 22-25 January 1995.
AB: Wine (1 ml) was acidified to pH 1-2 with 10% H2SO4 and extracted with 4 * 2 ml ethyl acetate. The aqueous layer was adjusted to pH 11 with 1M-NaOH, treated with 100 mul isobutyl chloroformate for 10

min and extracted with 4 * 2 ml diethyl ether. The aqueous layer was acidified to pH 1-2 with 10% H2SO4, saturated with NaCl and loaded onto a Chromosorb P SPE column. The retained N(OS)-isobutyloxycarbonyl amino-acids were eluted with diethyl ether and the eluate was evaporated to dryness. The residue was dervatized with 15 mul N-methyl-N(t-butyldimethylsilyl) trifluoroacetamide in 15 mul acetonitrile for 20 min at 60degC. The reaction mixture was analysed by dual column GC and GC-MS. For dual column GC, DB17 and DB5 columns (30 m * 0.25 mm i.d., 0.25 mum film thickness) were used with temperature programming and FID. For confirmation of peak identity by GC-MS, a column (25 m * 0.20 mm i.d.) coated with Ultra 2 (0.33 mum) was used with EIMS detection. Seventeen free amino-acids were identified in the four wines studied. Characteristic patterns for each wine were produced by simplifying the chromatograms to the corresponding amino-acid retention index spectra presented in a bar graphical form. Stepwise discriminant analysis of these profiles produced star symbols characteristic of each wine.

TI: Classification of tea samples by their chemical composition using discriminant analysis.
AU: Valera_P, Pablos_F, Gonzalez_AG
JN: Talanta, 1996, Vol.43, No.3, pp.415-419
IS: 0039-9140
DT: Article
NA: Dept. Anal. Chem., Univ. Seville, 41012 Seville, Spain
AB: Thirty samples of black or green tea were analysed for aqueous extract, total polyphenols and free amino-acids using gravimetric analysis, Folin-Ciocalteu reagent and ninhydrin, respectively. Caffeine, theobromine and theophylline were determined using reversed-phase HPLC with detection at 254 nm. This chemical analysis was combined with multivariate data interpretation; supervized pattern recognition methods were applied to the data obtained in order to establish discrimination rules to differentiate between the two classes of teas. Significant differences were detected between the groups for four of these descriptors: the amount of aqueous extract, polyphenols, free amino-acids and theophylline. A training set of ten black teas and five green teas was used to establish the classification rules by linear discriminant analysis with the software package CSS:STATISTICA (Statsoft). This procedure correctly classified the test set.

TI: Characterization of mineral waters by pattern recognition methods.
AU: Caselli_M, DeGiglio_A, Mangone_A, Traini_A
JN: Journal of the Science of Food and Agriculture, 1998, Vol.76, No.4, pp.533-536
IS: 0022-5142
DT: Article
NA: Dept. Chim., Univ. Bari, 70126 Bari, Italy
AB: Eighty three samples of mineral water from four different wells were analysed for 23 parameters (e.g. Li(I), Na(I), K(I), Ca(II), Mg(II), fluoride, nitrate, sulfate and chloride, CO2. Analyses were performed with use of atomic spectroscopy, ionic chromatography, spectrophotometry, titration, evaporation and conductimetry. Principle component analysis (PCA) was performed on the standardized data matrix with use of SCAN software. The PCA resulted in a feature reduction to two or three dimensions without major losses in information. Hierarchical and non-hierarchical procedures resulted in good separation of the data into four clusters. Samples from different wells were mostly assigned to different clusters.

TI: Gas chromatographic profiling and pattern recognition analysis of urinary organic acids for uterine myoma patients and cervical cancer patients.
AU: Kim_KR, Park_HG, Paik_MJ, Ryu_HS, Oh_KS, Myung_SW, Liebich_HM
JN: Journal of Chromatography, B: Biomedical Applications, 1998, Vol.712, No.1-2, pp.11-22
IS: 0378-4347
DT: Article
NA: Coll. Pharm., Sungkyunkwan Univ., Suwon 440-746, South Korea
AB: Urine samples were mixed with trans-cinnamic acid (internal standard) to 10 ppm, and a 1 ml portion was adjusted to pH 13 with 0.1M-NaOH and treated with methoxylamine hydrochloride (10 mg) at 60degC for 1 h. Portions (0.25 ml) were loaded onto a SAX tube, the tube rinsed with diethyl ether (0.25 ml) and the fatty acids were eluted with 0.1M-sodium sulfate (saturated with NaCl) and diethyl ether in sequence. The eluates were combined, the ether removed, the eluate was acidified with concentrated H2SO4 and saturated with NaCl. The acidified eluate was loaded onto a Chromosorb P tube, eluted with diethyl ether into a mixture of TEA (20 mul) and isooctane (40 mul). The excess ether was removed

under a stream of N2 and the residue was silylated with MTBSTFA (20 mul) at 60degC for 2 h. GC analysis was performed on the silylated mixture on a dual capillary system comprising fused-silica columns (30 m * 0.25 mm i.d.), packed with SE-54 and OV-17 bonded phases (0.25 mum) and FID detection. It was possible to identify 50 organic acids in the urine samples. When the GC profiles were put into bar graphical form, characteristic patterns were obtained for each average of benign and malignant tumour groups.

TI: Adaptation of linear discriminant analysis to second level pattern recognition classification.
AU: GonzalezArjona_D, Gonzalez_AG
JN: Analytica Chimica Acta, 1998, Vol.363, No.1, pp.89-95
IS: 0003-2670
DT: Article
NA: Dept. Phys. Chem., Univ. Seville, 41012 Seville, Spain
AB: A class modelling linear discriminant analysis (CMLDA) procedure is outlined and its application to second level pattern recognition is described. It was validated using three reference data sets for chemometrical methods testing (iris, olive oil and thyroid). The results were comparable with those obtained by the SIMCA and UNEQ class modelling techniques. The advantages, scope and limitations of CMLDA are discussed.

TI: CHEMICAL CHARACTERIZATION OF CRUDE-OIL RESIDUES FROM AN ARCTIC BEACH BY GC/MS AND GC/FID
AU: WANG_ZD, FINGAS_M, SERGY_G
NA: ENVIRONM CANADA, ETC, DIV EMERGENCIES SCI, 3439 RIVER RD, OTTAWA, ON K1A 0H3, CANADA
JN: ENVIRONMENTAL SCIENCE & TECHNOLOGY, 1995, Vol.29, No.10, pp.2622-2631
IS: 0013-936X
DT: Article
AB: A complete "total oil analysis method" suitable for monitoring chemical composition changes and studying the fate of 12-year-old weathered oil residues from an arctic beach is described. The characterizations not only are through analyses of individual aliphatic, aromatic, and biomarker compounds but also are through "pattern recognition" plots involving more than 100 important oil components and component groupings. The weathered percentages of residual oil in Baffin Island oil spill samples are quantified using C-29 and C-30 alpha beta-hopane in the "fresh" source oil as internal oil references. The values of the weathered percentages show excellent correlation to the total solvent-extractable materials (TSEM), total petroleum hydrocarbons (TPH), aliphatic, aromatic, and biomarker compound analysis results. Biodegradation is demonstrated to have played an important role in the degradation and removal of the residual oil. Twelve years after the spill, the composition changes due to weathering progress much more slowly, and this slower rate of change will continue under these arctic conditions.

TI: Study of the effects of weathering on the chemical composition of a light crude oil using GC/MS GC/FID
AU: Wang_ZD, Fingas_M
NA: ENVIRONM CANADA, ETC, DIV EMERGENCIES SCI, 3439 RIVER RD, OTTAWA, ON K1A 0H3, CANADA
JN: JOURNAL OF MICROCOLUMN SEPARATIONS, 1995, Vol.7, No.6, pp.617-639
IS: 1040-7685
DT: Article
AB: Quantitative information on the weathering of spilled oil is essential to a fuller understanding of the fate and behavior of oil in the environment. Such data is also useful for spill modeling. The key to acquiring data on oil weathering is the availability of precise and reliable chemical information.Exact quantitation of compounds in the oil can provide this crucial data. In this study, the effects of weathering on the chemical composition of a light crude oil, Alberta Sweet Mix Blend (ASMB), were thoroughly investigated using GC/FID and GC/MS. Complete compositional information on the ASMB oil at various degrees of evaporation (0-45%) was obtained, and the composition and concentration changes of key components and component groupings were quantitatively correlated to evaporative loss. Two opposing effects during evaporation-one is the loss of oil components due to evaporation, and another is build-up of oil components due to volume deduction-were examined. So-called "pattern recognition" plots involving more than 100 important individual oil components and component groupings were graphically depicted,

and these permitted deduction of a best set of values for quantitation of exposure to evaporative weathering. A "weathering index" concept is proposed. Relatively simple and very useful mathematical equations were derived which can be utilized to describe the weathering behavior of oil and to estimate the evaporation extent of oil.

TI: Classification of vegetable oils by FT-IR
AU: Dahlberg_DB, Lee_SM, Wenger_SJ, Vargo_JA
NA: LEBANON VALLEY COLL, DEPT CHEM, ANNVILLE, PA, 17003
JN: APPLIED SPECTROSCOPY, 1997, Vol.51, No.8, pp.1118-1124
IS: 0003-7028
DT: Article
AB: The Fourier transform infrared (FT-IR) spectra of 27 brands of 10 types of cooking oils and margarines were measured without temperature control. Attempts to predict the vegetable source and physical properties of these oils failed until wavelength selection and multiplicative signal correction (MSC) were applied to the FT-IR spectra. After pretreatment of the data, principal component analysis (PCA) was totally successful at oil identification, and partial least-squares (PLS) models were able to predict both the refractive indices [standard error of estimation (SEE) 0.0002] and the viscosities (SEE 0.52 cP) of the oils. These models were based predominately on the FT-IR detection of the cis and trans double-bond content of the oils, as well as small amounts of defining impurities in sesame oils. Efforts to use selected wavelengths to discriminate oil sources were only partially successful. These results show the potential utility of FT-IR in the fast detection of substitution or adulteration of products like cooking oils.

TI: MASS-SPECTROMETRIC PROFILING AND PATTERN-RECOGNITION
AU: TAS_AC, VANDERGREEF_J
NA: TNO, NUTR & FOOD RES INST, POB 360, ZEIST, NETHERLANDS
JN: MASS SPECTROMETRY REVIEWS, 1994, Vol.13, No.2, pp.155-181
IS: 0277-7037
DT: Review
AB: Much research in analytical chemistry is focused on the analysis of complex mixtures and trace analysis of compounds. In both cases, sensitive and, even more important, selective analytical tools for the determination of individual components are mandatory. However, investigation of complex and often diffuse (bio)chemical systems requires a different analytical approach aimed at the detection of a wide range of compounds rather than a selective analysis of compounds. This requirement is especially true if no prior knowledge exists of the kind of compounds involved in the properties of such systems. The analysis of macromolecular systems or complex matrices consisting of chemical components with strongly diverging physico-chemical properties can often hardly be performed through separation and determination of individual components. Examples of such matrices are bioreactor mixtures, microorganisms, cells of biological origin, body fluids, plant tissues, raw materials for food production, food products, dietary fiber, soil, humic substances, and fossil deposits (1-5). Especially the macromolecular part of such samples contains compounds that are highly involatile, with molecular weights far beyond the mass range of mass spectrometers. Mass spectrometry requires ions in vacuum. This process involves formation of ions from volatile compounds (generally EI and CI processes, or desorption of ions into the gas phase, for example by field desorption (FD) (6, 7) or charged aerosol droplets (thermospray, electrospray, ion spray) for nonvolatile compounds (8). Compounds of medium volatility and relatively high thermostability, usually in the medium molecular weight range, can be analyzed with thermal desorption techniques, such as direct probe mass spectrometry (DP-MS) (3). Apparently, these techniques are not appropriate for direct mass spectrometric analyses of the intact high-molecular-weight and heat-labile compounds of the complex matrices described above. For such samples, thermal degradation techniques (pyrolysis) can be performed to generate more volatile compounds of lower molecular weight that are amenable to mass spectrometric analysis (analytical pyrolysis). Therefore, many analytical profiling methods of (bio)macromolecular matrices are based on analysis of thermally-degraded matrices. Generally, application of such degradation techniques results in mixtures of high complexity. Therefore, advanced analytical equipment is needed to analyze the mixtures generated. Pyrolysis mass spectrometry (Py-MS) (4, 9-12), pyrolysis gas chromatography (Py-GC), and pyrolysis gas chromatography-mass spectrometry (Py-GC/MS) (5, 13, 14) are the techniques most widely applied in the field of analytical pyrolysis. They offer a powerful approach in that a great variety of compounds can be measured in one analysis and usually limited sample preparations are required. Generally, only small amounts of sample (typically 10-200 mug) are needed (1-5). Important applications

are fingerprinting for characterization and differentiation of polymers, biomaterials, and microorganisms using pattern recognition methods (4, 5, 15), and identification of individual compounds in pyrolysates for obtaining insight into the structure of the original bio)macromolecules (3, 4, 16-18). In the fingerprinting approach, differentiation is frequently based on statistical evaluation rather than on chemical interpretation of differentiating components (15). However, in most applications of this kind, at least global indications are obtained of relevant chemical compounds on which differentiation is based. Comparison of the techniques most widely used in analyses of this type, Py-GC and Py-MS, shows that dedicated Py-MS equipment offers a number of potential advantages (4, 9, 19), such as a high sample throughput, the potential of handling compounds differing to a large extent in polarity and molecular weight (4, 12), the suitability of MS data to computer handling for data analyses, good reproducibility, and a better long-term stability. However, for the identification of a large number of individual compounds in pyrolysates, for example, extended series of isomeric or homologous compounds, Py-GC/MS (5), Py-GC/(HR)MS (14, 17), and Py-GC/MSMS (20) are the techniques of choice, as long as volatile compounds are encountered. Serious limitation are the condensation of nonvolatile, relatively high-molecular-weight, and polar compounds on the inner wall of the glass linear and column and the relatively long analysis times.Recent developments in pyrolysate analysis embrace the application of tandem mass spectrometry (Py-MSMS) (18, 21-25), Py-GC/MSMS (20), and (off-line Py) fast atom bombardment tandem mass spectrometry (FAB/MSMS) (11). Py-MSMS is a powerful method for identifying molecular species directly in pyrolysates and has successfully been applied to the characterization of thermal decomposition products of polymers (21, 23), biopolymers (18, 24), bacteria (22), and algae (25). It is interesting to note that some recent reports show an apparent tendency to characterization of microorganisms using only well-defined chemical marker components (26, 27). The application of high-performance liquid chromatography (HPLC) and supercritical chromatography (SFC) to high-molecular-weight and polar compounds from pyrolysates has been reported recently (28, 29). In general, the approach requires off-line pyrolysis and pyrolysate derivatization preceding HPLC or SFC separation (29), but some initial results of on-line pyrolysis under SFC conditions with subsequent SFC analysis (Py-SFC) have been reported (28). Especially in SFC analysis of polar compounds, derivatization is required to reduce the polarity (30). Also, the direct monitoring of volatile compounds becomes increasingly important. Membrane introduction mass spectrometry (MIMS), applying hollow-fiber capillary membranes or membrane sheets, in combination with (tandem) mass spectrometry has recently been developed for the monitoring of mixtures of relatively volatile organic compounds in aqueous solutions, for gas analysis, and for in vivo studies of the formation of low-molecular-weight compounds. These techniques are applied to monitor small molecules in fermentation broths (fermentor control), to study microbial physiology, and for in vivo analysis of blood gases (31, 32). Flow injection/membrane introduction devices have been applied to bioreactor monitoring (33). Because data acquired from (bio)chemical matrices are mostly very complex, data analysis techniques are necessary for an effective evaluation of the analytical output. Already in an early stage of direct mixture analysis, in the late 1960s and the early 1970s, the first applications of data preprocessing and pattern recognition techniques were mentioned (4, 34). Data analysis techniques will be discussed in another section of this review. In this article, we will selectively review the use of direct mass spectrometric profiling for the analysis of complex (bio)chemical systems with an emphasis on applications in which pattern recognition techniques have been applied to the evaluation of the resulting profiles. Application of chromatographic separation techniques to mixture analysis is mentioned only occasionally in this review. The direct analysis of polymers by mass spectrometry has been reviewed separately in this journal (3) and, therefore, will not be discussed in this article.

TI: AUTOMATED GAS-CHROMATOGRAPHIC AMPHETAMINE PROFILING
AU: KARKKAINEN_M, SIPPOLA_E, PIKKARAINEN_AL, RAUTIO_T, HIMBERG_K
NA: NATL BUR INVEST, CRIME LAB, SF-00580 HELSINKI, FINLAND; VTT, BIOTECHNOL & FOOD RES, SF-02044 ESPOO, FINLAND
JN: FORENSIC SCIENCE INTERNATIONAL, 1994, Vol.69, No.1, pp.55-64
IS: 0379-0738
DT: Article
AB: A computerized procedure is presented for profiling and comparison of illicit amphetamine seizures. The system involves an optimized capillary gas chromatographic separation step followed by peak identification by linear retention indices. An individual amphetamine street sample can be recognized after up to 14 identified characteristic peaks are mathematically compared with the data in an intralaboratory database via an exponential comparison algorithm. The procedure enables rapid

screening: the GC analysis of an unknown sample and computer comparison against several hundred samples is performed in just over an hour.

TI: VOLATILE ORGANIC-COMPOUNDS IN THE BREATH OF PATIENTS WITH SCHIZOPHRENIA
AU: PHILLIPS_M, ERICKSON_GA, SABAS_M, SMITH_JP, GREENBERG_J
NA: ST VINCENTS MED CTR, DEPT MED, STATEN ISL, NY,10310; ST VINCENTS MED CTR, DEPT PSYCHIAT, STATEN ISL, NY,10310; INFOMETRIX INC, SEATTLE, WA; FIRST HILL DIAGNOST IMAGING CTR, SEATTLE,WA
JN: JOURNAL OF CLINICAL PATHOLOGY, 1995, Vol.48, No.5, pp.466-469
IS: 0021-9746
DT: Article
AB: Aims-To analyse the breath of patients with schizophrenia for the presence of abnormal volatile organic compounds. Methods-A case comparison study was performed in two community hospitals in Staten Island, New York. Twenty five patients with schizophrenia, 26 patients with other psychiatric disorders, and 38 normal controls were studied. Alveolar breath samples were collected from all participants, and volatile organic compounds in the breath were assayed by gas chromatography with mass spectroscopy. Differences in the distribution of volatile organic compounds between the three groups were compared by computerised pattern recognition analysis. Results-Forty eight different volatile organic compounds were observed in the breath samples. Three separate pattern recognition methods indicated an increased differentiation capability between the patients with schizophrenia and the other subjects. Pattern recognition category classification models using 11 of these volatile organic compounds identified the patients with schizophrenia with a sensitivity of 80.0% and a specificity of 61.9%. Volatile organic compounds in breath were not significantly affected by drug therapy, age, sex, smoking, diet, or race. Conclusions-Microanalysis of volatile organic compounds in breath combined with pattern recognition analysis of data may provide a new approach to the diagnosis and understanding of schizophrenia. The physiological basis of these findings is still speculative.

TI: IDENTIFICATION OF MYCOBACTERIUM-TUBERCULOSIS AND MYCOBACTERIUM-AVIUM COMPLEX DIRECTLY FROM SMEAR-POSITIVE SPUTUM SPECIMENS AND BACTEC 12B CULTURES BY HIGH-PERFORMANCE LIQUID-CHROMATOGRAPHY WITH FLUORESCENCE DETECTION AND COMPUTER-DRIVEN PATTERN-RECOGNITION MODELS
AU: JOST_KC, DUNBAR_DF, BARTH_SS, HEADLEY_VL, ELLIOTT_LB
NA: TEXAS DEPT HLTH,BUR LABS,DIV MICROBIOL SERV,1100 W 49TH ST,AUSTIN,TX,78756
JN: JOURNAL OF CLINICAL MICROBIOLOGY, 1995, Vol.33, No.5, pp.1270-1277
IS: 0095-1137
DT: Article
AB: A high-performance liquid chromatography method that utilized fluorescence detection (HPLC-FL) of mycolic acid 6,7-dimethoxycoumarin esters was developed to identify Mycobacterium tuberculosis (MTB) and M. avium complex (MAC) directly from fluorochrome stain smear-positive sputum specimens and young BACTEC 12B cultures. HPLC-FL chromatograms from a training set that included 202 smear-positive clinical sputum specimens and 343 mycobacterial cultures were used to construct a calibrated peak naming table and computer-based pattern recognition models for MTB and MAC. Pattern recognition model performance was measured with an evaluation see of samples that included 251 smear-positive clinical sputum specimens and 167 BACTEC 12B cultures. Evaluation sputum specimens were culture positive for MTB (n = 132) and MAC (n = 48). With evaluation sputa, the MTB and MAC models were 56.8 and 33.3% sensitive, respectively. Evaluation set BACTEC 12B cultures were culture positive for MTB (n = 97) and MAC (n = 53). The sensitivities of the MTB and MAC models for identification of BACTEC 12B cultures were 99.0 and 94.3%, respectively. The specificity of both models was 100% for both types of evaluation samples. The average times from BACTEC 12B inoculation to cell harvest mere 10.2 and 7.4 days for MTB and MAC, respectively. HPLC-FL can identify MTB and MAC in 1 day from many smear-positive sputa. Rapid and sensitive identification of MTB and MAC from young BACTEC 12B cultures was achieved.

TI: MULTIVARIATE-ANALYSIS FOR CLASSIFICATION OF COMMERCIAL ORANGE JUICE PRODUCTS BY VOLATILE CONSTITUENTS USING HEADSPACE GAS-CHROMATOGRAPHY
AU: SHAW_PE, MOSHONAS_MG, BUSLIG_BS
NA: USDA ARS, US CITRUS & SUBTROP PROD LAB, S ATLANTIC AREA, 600 AVE S NW, POB 1909, WINTER HAVEN, FL,33882; FLORIDA DEPT CITRUS, WINTER HAVEN, FL, 33882
JN: ACS SYMPOSIUM SERIES, 1995, Vol.596, pp.33-47
IS: 0097-6156
DT: Review
AB: Analyses of fresh and processed orange juices by headspace gas chromatography afforded quantities of up to 40 volatile components in each juice type. Many of these components are known to influence citrus juice flavor. Multivariate analysis of the quantitative date with a computer pattern recognition program classified the various juice samples according to processing conditions. The graphically displayed classifications corresponded to expected flavor quality. These results can potentially help processors determine product quality without sensory evaluation measurements, and suggest changes in processing conditions to improve flavor of processed products.

TI: Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques
AU: Vogels_JTWE, Terwel_L, Tas_AC, vandenBerg_F, Dukel_F, vanderGreef_J
NA: TNO, NUTR & FOOD RES INST, CTR STRUCT ELUCIDAT & INSTRUMENTAL ANAL, POB 360, 3700 AJ ZEIST, NETHERLANDS; INSPECTORATE HLTH PROTECT, 3001 KB ROTTERDAM, NETHERLANDS; TNO, NUTR & FOOD RES INST, DEPT ANAL, 3700 AJ ZEIST, NETHERLANDS
JN: JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY, 1996, Vol.44, No.1, pp.175-180
IS: 0021-8561
DT: Article
AB: This paper describes the application of proton NMR spectroscopy as a screening tool for the determination of the authenticity of orange juices. Principal component and discriminant analyses were used to discriminate between authentic and nonauthentic (suspect) samples. In one set of profiles, additions of sucrose, beet medium invert sugar, sodium benzoate could easily be detected. In another set of data, K-nearest neighbor classification based on the principal component scores was used to correctly classify at least 94% of samples known to deviate from authentic samples when measured with analytical techniques such as high pressure liquid chromatography and atomic absorption spectroscopy. Principal component loading plots and factor spectra were an effective tool in the interpretation of the differences between the profiles.

TI: Pattern recognition applied to gas chromatographic profiles of volatile components in three tea categories
AU: Togari_N, Kobayashi_A, Aishima_T
NA: GIFU WOMENS UNIV, 80 TAROMARU, GIFU 50125, JAPAN; OCHANOMIZU UNIV, BUNKYO KU, TOKYO 112 ,JAPAN; KIKKOMAN FOODS INC, NODA, CHIBA 278, JAPAN
JN: FOOD RESEARCH INTERNATIONAL, 1995, Vol.28, No.5, pp.495-502
IS: 0963-9969
DT: Article
AB: Volatile components in unfermented green tea, semi-fermented Oolong tea and fully fermented black tea were analyzed by gas chromatography (GC) and gas chromatography-mass spectrometry (CC-MS). For differentiating, three tea categories based on their volatile components, unsupervised and supervised pattern recognition techniques were applied to the resulting GC data. Three distinct clusters each corresponding to green tea, Oolong tea and black tea were observed in the dendrogram and the principal component (PC) score plot. However, a subcluster of Oolong tea was observed in the vicinity of black tea cluster in both the dendrogram and the PC plot. The first and second PC corresponded to the fermentation products and aroma components originally contained in tea leaves, respectively. Both the partial least squares (PLS) analysis and linear discriminant analysis correctly differentiated tea samples into the three categories. (E)-2-Hexenal, the major fermentation product from unsaturated fatty acids, was the most efficient for the discrimination. Although three teas are produced from the same plant species, pattern recognition clarified the existence of the apparent quality difference among their volatile component profiles.

TI: Prediction of gas chromatographic retention indices of alkylbenzenes.
AU: Sutter_JM, Peterson_TA, Jurs_PC
NA: PENN STATE UNIV, DEPT CHEM, DAVEY LAB 152, UNIVERSITY PK, PA, 16802.
JN: ANALYTICA CHIMICA ACTA, 1997, Vol.342, No.2-3, pp.113-122
IS: 0003-2670
DT: Article
AB: The retention indices (RIs) of a set of alkylbenzenes on a polar gas chromatographic column are predicted directly from their molecular structures. Numerical descriptors are calculated based on the structure of a group of 150 alkylbenzenes. The descriptors are of three types: topological, geometric, and electronic. Statistical methods are employed to find an informative subset of these descriptors that can accurately predict the gas chromatographic RIs. The Automated Data Analysis and pattern Recognition Toolkit (ADAPT) software system is used to construct a large pool of structurally derived numerical descriptors which are used to build quantitative structure-retention relationships (QSRRs). Multiple linear regression analysis and computational neural networks are used to map the descriptors to the RIs.

TI: Characterisation of citrus by chromatographic analysis of flavonoids
AU: Robards_K, Li_X, Antolovich_M, Boyd_S
NA: CHARLES STURT UNIV, SCH SCI & TECHNOL, POB 588, WAGGA WAGGA, NSW 2678, AUSTRALIA; CHARLES STURT UNIV, ENVIRONM & ANALYT LABS, WAGGA WAGGA, NSW 2678, AUSTRALIA
JN: JOURNAL OF THE SCIENCE OF FOOD AND AGRICULTURE, 1997, Vol.75, No.1, pp.87-101
IS: 0022-5142
DT: Article
AB: Flavonoid glycosides in citrus were characterised by high-performance liquid chromatography using both ultraviolet and fluorescence detection. The effects of sample preparation on the chromatographic profiles are reported. Key variables in the profiles useful as chemotaxonomic markers were identified with the aid of pattern recognition, which was also used to create sample categories. LC-MS data are presented and the advantages of mass spectrometric detection are demonstrated.

TI: Chemometric aspects of sugar profiles in fruit juices using HPLC and GC
AU: Yoon_JH, Kim_K, Lee_DS
NA: SEOUL WOMENS UNIV, DEPT CHEM, SEOUL 139774, SOUTH KOREA; KOREA ADV INST SCI & TECHNOL, SAM,AES, ADV ANAL CTR SIMS, SEOUL 130650, SOUTH KOREA
JN: BULLETIN OF THE KOREAN CHEMICAL SOCIETY, 1997, Vol.18, No.7, pp.695-702
IS: 0253-2964
DT: Article
AB: The objective of this work is to determine the sugar profiles in commercial fruit juices, and to obtain chemometric characteristics. Sugar compositions of fruit juices were determined by HPLC-RID and GC-FID via methoxymation and trimethylsilylation with BSTFA. The appearance of multiple peaks in GC analysis for carbohydrates was disadvantageous as described in earlier literatures. Fructose, glucose, and sucrose were major carbohydrates in most fruit juices. Glucose/fructose ratios obtained by GC were lower than those by HPLC. Orange juices are similar to pineapple juices in the sugar profiles. However, grape juices are characterized by its lower or no detectable sucrose content. In addition, it was also found that unsweeten juices contained considerable level of sucrose. Chemometric technique such as principal components analysis was applied to provide an overview of the distinguishability of fruit juices based on HPLC or GC data. Principal components plot showed that different fruit juices grouped into distinct cluster. Principal components analysis was very useful in fruit juices industry for many aspects such as pattern recognition, detection of adulterants, and quality evaluation.

TI: Use of high-performance liquid chromatographic chemometric techniques to differentiate apple juices clarified by microfiltration and ultrafiltration
AU: BlancoGomis_D, FernandezRubio_P, GutierrezAlvarez_MD, MangasAlonso_JJ
NA: UNIV OVIEDO, FAC QUIM, DEPT QUIM FIS & ANALIT, E-33006 OVIEDO, SPAIN; CTR INVEST APLICADA & TECNOL AGROALIMENTARIA, E-33300 VILLAVICIOSA, SPAIN
JN: ANALYST, 1998, Vol.123, No.1, pp.125-129
IS: 0003-2654
DT: Article

AB: HPLC of amino acids and riboflavin in apple juices clarified by means of cross-flow membrane technology was used to characterise the juices. The chromatographic information was subjected to pattern recognition methods such as principal component analysis, K-nearest neighbour (KNN), linear discriminant analysis, Bayes analysis, soft independent modelling of class analogy and partial least squares.

TI: Characterization of fatty acids composition in vegetable oils by gas chromatography and chemometrics
AU: Lee_DS, Noh_BS, Bae_SY, Kim_K
NA: SEOUL WOMENS UNIV, DEPT CHEM, SEOUL 139774, SOUTH KOREA; SEOUL WOMENS UNIV, DEPT FOOD SCI, SEOUL 139774, SOUTH KOREA
JN: ANALYTICA CHIMICA ACTA, 1998, Vol.358, No.2, pp.163-175
IS: 0003-2670
DT: Article
AB: Principal component analysis and discriminant analysis were applied to gas chromatographic data for fatty acids composition of commercial edible vegetable oils including sesame, perilla, soybean, corn germ, canola, rapeseed, olive and coconut oils. Principal components plot showed that eight different vegetable oils are clustered in distinct groups; each group could be distinguished clearly. Discriminant analysis was employed to assign unknown samples into one of two groups. Principal component analysis or discriminant analysis is very useful for many aspects of vegetable oils industry, including pattern recognition or primary evaluation of category similarity, detection of adulterants, and quality control.

TI: Gas chromatography mass spectral analysis of roots of Echinacea species and classification by multivariate data analysis
AU: Lienert_D, Anklam_E, Panne_U
NA: COMMISS EUROPEAN COMMUNITIES, INST ENVIRONM, JOINT RES CTR, I-21020 ISPRA, ITALY; INST HYDROCHEM, D-81377 MUNICH, GERMANY
JN: PHYTOCHEMICAL ANALYSIS, 1998, Vol.9, No.2, pp.88-98
IS: 0958-0344
DT: Article
AB: An analytical method based on gas chromatography-mass spectral (GC-MS) analysis was developed as a fast screening tool in order to verify the authenticity of extracts of roots from different species of Echinacea, namely E. angustifolia, E. pallida and E. purpurea. Various extraction methods, i.e. soxhlet extraction, supercritical fluid extraction and maceration with three different solvents, were applied and the extracts were analysed by GC-MS, The chromatograms were evaluated by multivariate data analysis, i.e. cluster analysis, principal component analysis;and discriminant analysis, in order to reveal if a classification into the three main species of Echinacea was possible using the Information obtained, GC-MS analysis of the extracts of Echinacea, together with multivariate data analysis, displayed substantial classification power since a good separation of the three different species was achieved, This analytical approach was not only suitable for classification but was also sufficiently robust such that no distortion of root samples by ageing occurred.

TI: Authentication of green coffee varieties according to their sterolic profile
AU: Carrera_F, LeonCamacho_M, Pablos_F, Gonzalez_AG
NA: UNIV SEVILLA,DEPT ANALYT CHEM, E-41012 SEVILLE, SPAIN; CSIC, INST GRASA, E-41012 SEVILLE, SPAIN
JN: ANALYTICA CHIMICA ACTA, 1998, Vol.370, No.2-3, pp.131-139
IS: 0003-2670
DT: Article
AB: Sterols of 31 samples of green coffee beans of the arabica and robusta varieties have been analysed by gas chromatography flame ionization detector. The lipids were Soxhlet extracted from ground coffee beans into hexane. The extract was evaporated and the residue was treated with 0.2% of Sa-cholestane-SP-ol (internal standard). The lipids were saponified and the sterolic fraction of the unsaponifiable part was separated by thin layer chromatography. Sterols were treated with a silylating reagent to convert them into trimethyl silyl derivatives. This analysis was performed on a column (30 mx0.32 mm i.d.) coated with a bonded stationary phase HP-5 (5% diphenyl-95% methylpolysiloxane; 0.2 mu m) with hydrogen (0.7 ml min(-1)) as a carrier gas, isothermal temperature at 265 degrees C and flame ionization detector. By using the sterols as chemical descriptors pattern recognition techniques were applied to differentiate

between arabica and robusta green coffee varieties. Delta(5)avenasterol and sitostanol were found to be the most differentiating variables.

TI: Differentiation of soy sauce types by HPLC profile pattern recognition - Isolation of novel isoflavones
AU: Kinoshita_E, Ozawa_Y, Aishima_T
NA: KIKKOMAN FOODS INC, DIV RES & DEV, 399 NODA, CHIBA 2780037, JAPAN
JN: ADVANCES IN EXPERIMENTAL MEDICINE AND BIOLOGY, 1998, Vol.439, pp.117-129
IS: 0065-2598
DT: Article
AB: Nonvolatile minor components in various brands of Japanese fermented soy sauce were analyzed by gradient RP-HPLC and monitored at 280 nm. Chemometric pattern recognition techniques, such as cluster analysis, linear discriminant analysis (LDA), LDA using genetic algorithm (GA-LDA) and soft independent modelling of class analogy (SIMCA), succeeded in differentiating the resulting HPLC profiles according to soy sauce brands. Three components playing key roles in the differentiation were isolated by preparative HPLC and purified by gel-filtration chromatography, or simply repeated preparative HPLC. FAB-MS, H-1-, C-13-NMR and IR spectra suggested that these three components having molecular weights of 386, 402 and 418 were isoflavone derivatives. By applying HMBC spectral analysis, these isoflavones were identified as conjugated ethers of tartaric acid with daidzein, genistein and 8-hydroxygenistein. These new isoflavone derivatives are produced by some strains of Aspergillus fungi.

4. Neural network and staistical pattern recognition techniques for quality assurrance

TI: The peak tracking role of the prima method in liquid chromatography.
AU: Nemeth_Z
JN: ACH MODELS IN CHEMISTRY, 1994, Vol.131, No.6, pp.835-845
IS: 1217-8969
DT: Article
NA: Dept. Chem., Univ. Forestry and Wood Sci., 9401 Sopron, Hungary
AB: A new pattern recognition method based on a modification of the identification function of the Pattern Recognition by Independent Multicategory Analysis (PRIMA) method is described for peak tracking in the LC analysis of unknown samples. The method was applied to the separation of a mixture containing benzoic acid, salicylic acid, phthalamidic acid, phthalic acid, anthranilic acid, and isatoic anhydride. Using aqueous 15-30% methanol of pH 1.5-6.5 (mobile phase) and an ODS- Hypersil column (cf. Szokoli et al., Chromatographia, 1990, 29, 265). The method was independent of material quality and enabled optimal separation conditions to be selected.

TI: Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods.
AU: Malmquist_G, Danielsson_R
JN: Journal of Chromatography, A, 1994, Vol.687, No.1, pp.71-88
IS: 0021-9673
DT: Article
NA: Inst. Chem., Dept. Anal. Chem., Uppsala Univ., 751 21 Uppsala, Sweden
CO: Presented at the 6th International Symposium on High Performance Capillary Electrophoresis, held in San Diego, CA, USA, 31 Jan-3 Feb 1994.
AB: For chromatogram fingerprinting, pattern recognition techniques based on principal components analysis require that chromatographic variations can be distinguished from true variations due to sample composition. A preprocessing procedure for chromatographic data is described, which facilitates the characterization of a set of reference chromatograms by principal components analysis. The procedure was used to align a sample chromatogram towards a target chromatogram in order to compensate for small shifts in retention time (not due to different sample components), common variations in peak areas (not due to sample composition) and variations in level and slope of the baseline. The effects of the procedure on the principal components analysis was demonstrated for a set of chromatographic profiles intended for peptide mapping.

TI: Evaluation procedures for reversed-phase high-performance liquid chromatographic columns in the analysis of strongly basic compounds using principal component analysis for data assessment.
AU: Brereton_RG, McCalley_DV
JN: Analyst (Cambridge, U. K.), 1998, Vol.123, No.6, pp.1175-1185
IS: 0003-2654
DT: Article
NA: School Chem., Univ. Bristol, Bristol BS8 1TS, UK
CO: Presented at the XXX Colloquim Spectroscopicum Internationale (CSI), held in Melbourne, Australia, September 21-26, 1997
AB: The performance of eight silica-based reversed-phase HPLC columns was evaluated using chemometric pattern recognition. Tests were performed at pH 3 and 7 using ten compounds, three mobile phase modifiers and four column evaluation parameters. Principal components analysis was performed on the data subsets. The results are discussed and recommendations are given for conducting tests to assess the quality of a column for the analysis of basic compounds.

TI: Assessment of chromatographic peak purity by means of artificial neural networks.
AU: Hu_YH, Zhou_GW, Kang_JH, Du_YX, Huang_F, Ge_JH
JN: Journal of Chromatography, A, 1996, Vol.734, No.2, pp.259-270
IS: 0021-9673
DT: Article
NA: Dept. Anal. Chem., China Pharm. Univ., Nanjing 210009, China
CO: Presented at the Seventh Symposium on Handling of Environmental and Biological Samples in Chromatography, held in Lund, Sweden, 8-10 May, 1995
AB: The artificial neural network method for the assessment of peak purity used a non-linear transformation function with a back-propagation algorithm to describe and predict the chromatographic data and the Mann-Whitney U-test to determine peak purity. The computer programs were written in Fortran 77 and run on a 486 computer under MS-DOS 6.2. The method was tested with simulated and experimental data for pure and mixed samples. The results obtained for the simulated data sets were compared to those obtained by principal component analysis. A prior knowledge of the impurity was not required.

TI: Artificial neural networks aided deconvolving overlapped peaks in chromatograms.
AU: Miao_HJ, Yu_MH, Hu_SX
JN: Journal of Chromatography, A, 1996, Vol.749, No.1-2, pp.5-11
IS: 0021-9673
DT: Article
NA: Lab. Intelligent Information Eng., Dept. Chem. Eng., Zhejiang Univ., Hangzhou 310027, China
CO: Presented at the 7th International Symposium on High Performance Capillary Electrophoresis, held in Wuerzburg, Germany, January 29-February 2, 1995
AB: A novel method for the deconvolution of overlapped chromatographic peaks was developed. Parameters characteristic of the shape of asymmetric gaussian peaks were identified and used with a multilayered perceptron network to determine individual peak areas for unresolved components. The artificial neural network method was shown to be more precise than the vertical line splitting and curve-fitting methods. Furthermore it required less computer time than conventional methods.

TI: NEURAL-NETWORK ASSISTED RAPID SCREENING OF LARGE INFRARED SPECTRAL DATABASES
AU: KLAWUN_C, WILKINS_CL
NA: UNIV CALIF RIVERSIDE, DEPT CHEM, RIVERSIDE, CA, 92521
JN: ANALYTICAL CHEMISTRY, 1995, Vol.67, No.2, pp.374-378
IS: 0003-2700
DT: Article
AB: A new method of prefiltering applicable to spectral database searches has been developed. It employs a backpropagation neural network to classify 609 matrix isolation FT-IR spectra with respect to the presence or absence of 35 functional groups, sewing as sortable bit string keys to the spectral library. These bit strings or integers are used to construct a binary search tree for 100% successful fast spectral retrieval. Compared with a sequential library search, this method yields a more than 25-fold increase in search speed and could easily be adapted to handle other types of spectral information. For larger databases, this advantage is expected to increase by a factor of n(0.5) relative to the number of spectra.

Additionally, the backpropagation neural network training has been modified with an extended version of the flashcard algorithm for 100% successful training of 2651 matrix isolation FT-IR spectra.

TI: Large artificial neural networks applied to the prediction of retention indices of acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones and ethers
AU: Yan_AZ, Zhang_RS, Liu_MC, Hu_ZD, Hooper_MA, Zhao_ZF
NA: LANZHOU UNIV, DEPT CHEM, LANZHOU 730000, PEOPLES R CHINA; MONASH UNIV, GIPPSLAND CTR ENVIRONM SCI, MONASH, AUSTRALIA
JN: COMPUTERS & CHEMISTRY, 1998, Vol.22, No.5, pp.405-412
IS: 0097-8485
DT: Article
AB: Artificial neural networks (ANN) with extended delta-bar-delta (EDBD) back propogation learning algorithms were used to predict the retention indices of 184 organic compounds. These compounds include acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones and ethers. The network's architecture and parameters were optimized to give maximum performance. The best network is 2-6-1, the optimum learning epoch is 2000. In the process of the study, cross-validation and leave-20%-out were used. The results show that the prediction performance of ANN operating with such non-linear systems is remarkably good.

TI: Carboxylic acids: prediction of retention data from chromatographic and electrophoretic behaviours
AU: Bruzzoniti_MC, Mentasti_E, Sarzanini_C
NA: UNIV TURIN, DEPT ANALYT CHEM, VIA P GIURIA 5, I-10125 TURIN; ITALY; UNIV TURIN, DEPT ANALYT CHEM, I-10125 TURIN, ITALY
JN: JOURNAL OF CHROMATOGRAPHY B, 1998, Vol.717, No.1-2, pp.3-25
IS: 0378-4347
DT: Review
AB: A review of the main results reached in the prediction of retention data of carboxylic acids, inferred by their chromatographic and electrophoretic behaviour, is presented. Attention has been focused on the main separation methods used in carboxylic acids analysis, that is ion-exclusion, anion-exchange, reversed-phase (RP) liquid chromatography and capillary electrophoresis. Papers proposing mechanistic models as well as chemometric and multilayer feed-forward neural network analysis of ion chromatography (IC) and RP chromatographic retention data were reviewed. Principal component analysis, PCA, sequential simplex method and simultaneous modelling of response surfaces through simple nonlinear models (not related to equilibria involved in retention) have been considered. Computer simulations for the prediction of retention data have also been discussed. A quick overlook on the prediction of capacity factors of analytes by less common determination methods such as thin-layer, gas chromatography and supercritical fluid chromatography has also been done.

TI: Prediction of flame ionization detector response factors using an artificial neural network
AU: JalaliHeravi_M, Fatemi_MH
NA: SHARIF UNIV TECHNOL, DEPT CHEM ,POB 11365-9516, TEHRAN, IRAN
JN: JOURNAL OF CHROMATOGRAPHY A, 1998, Vol.825, No.2, pp.161-169
IS: 0021-9673
DT: Article
AB: An artificial neural network (ANN) was successfully developed for the modeling of flame ionization detector response factors. The generated ANN was evaluated and applied for the prediction of response factors of several varieties of organic compounds. The results obtained using neural network were compared with different sets of experimental values as well as with those obtained using multiple linear regression technique. Comparison of neural network standard error of prediction values with those obtained using regression equations shows the superiority of ANNs over that of regression models. Calculations of Dietz response factor for two different prediction sets show that an ANN has a good predictive power.

TI: Application of an artificial neural network in chromatography-retention behavior prediction and pattern recognition
AU: Zhao_RH, Yue_BF, Ni_JY, Zhou_HF, Zhang_YK
NA: CHINESE ACAD SCI, DALIAN INST CHEM PHYS, NATL CHROMATOG R & A CTR, 161 ZHONGSHAN RD, DALIAN 116011, PEOPLES R CHINA
JN: CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, 1999, Vol.45, No.1-2, pp.163-170
IS: 0169-7439
DT: Article
AB: Layered feed-forward neural networks are powerful tools particularly suitable for the analysis of nonlinear multivariate data. In this paper, an artificial neural network using improved error back-propagation algorithm has been applied to solve problems in the field of chromatography. In this paper, an artificial neural network has been used in the following two applications: (1) To model retention behavior of 32 solutes in a methanol-tetrathydrofuran-water system and 49 solutes in methanol-acetonitrile-water system as a function of mobile phase compositions in high performance liquid chromatography. The correlation coefficients between the calculated and the experimental capacity factors were all larger than 0.98 for each solute in both the training set and the predicting set. The average deviation for all data points was 8.74% for the tetrahydrofuran-containing system and 7.33% for the acetonitrile-containing system. 2). To classify and predict two groups of different liver and bile diseases using bile acid data analyzed by reversed-phase high performance liquid chromatography (RP-HPLC). The first group includes three classes: healthy persons, choledocholithiasis patients and cholecystolithiasis patients; the total consistent rate of classification was 87%. The second group includes six classes: healthy persons, pancreas cancer patients, hepatoportal high pressure patients, cholelithiasis patients, cholangietic jaundice patients and hepatonecrosis patients; the total consistent rate of classification was 83%. It was shown that artificial neural network possesses considerable potential for retention prediction and pattern recognition based on chromatographic data.

TI: Application of the pattern-recognition method for modelling expert estimation of chromatogram quality
AU: Pirogov_AV, Platonov_MM, Pletnev_IV, Obrezkov_ON, Shpigun_OA
NA: MOSCOW MV LOMONOSOV STATE UNIV, DIV ANALYT CHEM, DEPT CHEM, GSP-3, MOSCOW 199899, RUSSIA
JN: ANALYTICA CHIMICA ACTA, 1998, Vol.369, No.1-2, pp.47-56
IS: 0003-2670
DT: Article
AB: A method of nonparametric regression, namely the method of k-nearest neighbours, was applied in chromatography for creating an estimation criterion of chromatogram quality. A training set was constructed using expert estimation of a number of model and real chromatograms, and the method was learned to estimate chromatograms as a single image. The parameters for chromatogram estimation were selected, and their variance weights were determined. The proposed approach is compared with the existing criteria, and its advantages are shown. The method was successfully applied for searching for the optimum conditions for the ion-chromatographic determination of inorganic anions.

TI: Alternate path reasoning in intelligent instrument fault diagnosis for gas chromatography
AU: Adair_KL, Hruska_SI, Elling_JW
JN: IEEE International Joint Symposia on Intelligence and Systems, 1996, pp.89-96
PU: IEEE, Los Alamitos, CA, USA
DT: Conference Paper
NA: Florida State Univ, Tallahassee, FL, USA
CF: Proceedings of the 1996 IEEE International Joint Symposia on Intelligence and Systems, Rockville, MD, USA, Nov 4-5 1996, (Conf. code 45792)
AB: Intelligent instrument fault diagnosis is addressed in this paper using expert networks, a hybrid technique which blends traditional rule-based expert systems with neural network style training. One of the most difficult aspects of instrument fault diagnosis is developing an appropriate rule base for the expert network. Beginning with an initial set of rules given by experts, a more accurate representation of the reasoning process can be found using example data. A methodology for determining alternate paths of reasoning and incorporating them into the expert network is presented. Our technique presupposes interaction and cooperation with the expert, and is intended to be used with the assistance of the expert to incorporate knowledge discovered from the data into the intelligent diagnosis tool. Tests of this

methodology are conducted within the problem domain of fault diagnosis for gas chromatography. Performance statistics indicate the efficacy of automating the introduction of alternate path reasoning into the diagnostic reasoning system. (Author abstract) 14 Refs.

TI: Mechanism of degradation of poly(vinyl butyral) using thermogravimetry/Fourier transform infrared spectrometry
AU: Liau_LCK, Yang_TCK, Viswanath_DS
JN: Polymer Engineering and Science, Mid-Nov 1996, Vol.36, No.21, pp.2589-2600
PU: Soc of Plastics Engineers, Brookfield, CT, USA
IS: 0032-3888
DT: Article
NA: Univ of Missouri, Columbia, MO, USA
AB: A TG/FTIR system was used to identify the products of thermal oxidative degradation of PVB, and also to elucidate the mechanism of degradation. This technique is useful in the kinetic analysis of fast reactions such as polymer degradation, unlike the use of a TG/GC/FTIR system, in which long retention times are needed to separate the products. A computer resolution method based on a pattern recognition technique is proposed to resolve the dynamic mixture IR spectra of the degradation products. A four-component synthetic mixture was used to evaluate the performance of the resolution algorithm and was found to be accurate within 5%. The method was then applied to PVB degradation. The dynamic information of PVB thermal oxidative degradation obtained by resolving the mixture IR spectra was used to elucidate the reaction mechanism and to determine the kinetic parameters. Results showed that PVB degradation in air took place at a temperature 50 K lower and the overall activation energy dropped from 338 kJ/mole (in nitrogen) to 200 kJ/mole (in air) compared with the degradation in a nitrogen atmosphere. (Author abstract) 25 Refs.

Further References from ISI (Science Citation Index)

1. TI: Neural networks for chromatographic peak classification - a preliminary study.
   AU: Rowe R.C., Mulley V.J., Hughes J.C., Nabney I.T., Debenham R.M.
   JN: LC GC-Magazine of Separation Science, 1994, Vol.12, No.9, p.690 et seq.
   AB: The authors applied neural networks, specifically the multilayer perceptron network, to classify peak shapes in chromatography. They compared a trained, optimized neural network with a human expert, both of which classified 396 individual peak profiles. Although both exhibited an 85% overall success rate, the neural network performed the task in 5.6 s, compared with 8 h for the human expert. In addition, the neural network showed complete objectivity.

2. TI: Global recognition of chromatographic profiles of complex matrices using mass spectrometry: Applications to flavorings and fragrances
   AU: Fellous_R, Gavini_P
   JN: ANALYSIS, 1997, Vol.25, No.6, pp.M24-M28

3. TI: Automation of gas chromatography instruments .1. Automated peak identification in the chromatograms of standard test mixtures
   AU: Du H, Stillman M.J.
   JN: ANALYTICA CHIMICA ACTA, 1997, Vol.354, No.1-3, pp.65-76
   AB: The evaluation of the chromatogram recorded during the gas chromatographic analysis of a standard test mixture is of importance in the assessment of the performance of the underlying GC system. Automated evaluation is essential in the development of software that can be used in systems that can perform totally automated analysis. In this study, an algorithm for automated peak identification in the chromatogram of the standard test mixture has been developed. In the proposed method, a specified test mixture is analyzed and the resultant chromatogram is stored. Peak identification for this experimentally determined chromatogram, which it is assumed will incorporate all recurring problems that might exist for all analyses, is accomplished by matching the measured peaks with peaks in the reference chromatogram The proposed algorithm is partly based on information theory and partly on chemical knowledge about the standard test mixture. The algorithm may be used to identify highly distorted peaks, peak retention time shifts (both positive and negative with respect to the reference chromatogram), missing peaks, and irregular peak retention times (shifts in both directions for different peaks). The program gives the user the ability to display the chromatogram and calculate four peak parameters from the experimental results of the test mixture under consideration: (i) peak retention time, (ii) the peak area counts, (iii) the peak width, and (iv) the asymmetry factor of the peak. Satisfactory performance is obtained using a set of simulated data that exhibit a range of problems that commonly occur in gas chromatograms as analyses proceed over time. (C) 1997 Elsevier Science B.V.

Miscellaneous References from Web Sites

Bruce L., G. Schmidt, (1994). Hydrocarbon Fingerprinting for Application in Forensic Geology: Review with Case Studies: AAPG Bulletin, V. 78, No.11

Carlson, P.;E. Reimnitz, (1990). Characterization of samples sites along the oil spill trajectory in Prince William Sound and the Gulf of Alaska. U.S. Geological Survey Open-File Report 90-39A. 23 p. in U.S. Geological Survey Open-File Report 90-39 (P. Carlson & E. Reimnitz, eds).

Christensen L., T. Larson, (1993). Method for determining the age of diesel oil spills in the soil: Ground Water Monitoring and Remediation, V. 13, No. 4

Curiale J., D. Cameron, D. Davis, (1985). Biological marker distribution and significance in oils and rocks of the Monterey Formation, California. Geochem Cosmochim. Act.

Kvenvolden K., J. Rapp, F. Hosteller, (1991). Tracking hydrocarbons from the Exxon Valdez oil spill in beach, shallow-water, and deep-water sediment of Prince William Sound, Alaska. In U.S. Geological Survey Open-File Report 91-631 (P. Carlson, ed.), 69-98.

Kvenvolden K., F. Hostettler, J. Rapp, P. Carlson, (1993). Hydrocarbons in oil residues on beaches of islands of Prince William Sound, Alaska. Marine Pollution Bulletin, 26(1) 1993.

Levorsen A., (1967) Geology of petroleum: San Francisco, W. H. Freeman, 176-231

Magoon L., C. Issacs, (1983) Chemical characteristics of same crude oils from the Santa Maria Basin, California. In Petroleum Generation and Occurrence in the Miocene Monterey Formation, California (C. Issacs & R. Garrison, eds), 201-211. Pacific Section, SEPM

McMillen S., N. Gray, J. Kewrr, A. Requejo, T. McDonald, G. Douglas, (1995). Assessing bioremediation in crudes and sludges. In monitoring and verification of bioremediation (R. Hinchee, G. Douglas, S. Ong, eds.), 1-9, Battelle Press, Columbus, OH

Rapp J., F. Hostlettler, K. Kvenvolden, (1990) Comparison of Exxon Valdez oil with extractable material from deep-water bottom sediment in Prince William Sound and the Gulf of Alaska. In U.S. Gelogical Survey Open-File Report 90-39 (P. Carlson & E. Reimnitz, eds.). Bottom sediment along oil spill trajectory in Prince William Sound and along Kenai Peninsula, Alaska.

Senn, R.B. and Johnson, M.S., "Interpretation of Gas Chromatographic Data in Subsurface Hydrocarbon Investigations", Ground Water Monitoring Review, p. 58-63, Winter 1987.

Shaw D., B. Baker, (1978). Hydrocarbons in the marine environment of Port Valdez, Alaska. Environ. Sci. Tech. 12, 1200-1205

Shaw D., T. Hogan, D. McIntosh, (1985). Hydrocarbons in sediments of Port Valdez, Alaska: consenquences of five years permitted discharge. Estur. Coast. Shelf Sci. 21, 131-144

Thomas, D.H. and Delfino, J.J., "A Gas Chromatographic/Chemical Indicator Approach to Assessing Ground Water Contamination by Petroleum Products", Ground Water Monitoring Review, p. 90-100, Fall 1991.

Watts G., (1989). Groundwater monitoring parameters and pollution sources. Third Edition Florida Department of Environmental Regulation. Tallahassee, Florida, pp. 65-70

CCME, "Guidance Manual on Sampling, Analysis and Data Management for Contaminated Sites", Volumes I (EPC-NCS62E) and II (EPC-NCS66E), Winnipeg, December 1993.

APPENDIX B

The Source Identification

Feasibility Study

-

The Questionnaire

Staffordshire University - School of Computing

-

Centre for Intelligent Environmental Systems

PATTERN MATCHING FOR THE SOURCING OF OIL SLICKS

As part of a current project at the Centre for Intelligent Environmental Systems at Staffordshire University, we are investigating aspects of *pattern matching* and *pattern recognition* for the sourcing of oil pollution in rivers. A *gas chromatogram* (GC) is a graph produced to analyse the different compounds which make up an oil (it is often referred to as the 'fingerprint' of the oil). We believe that patterns found in small sections of the GCs of oil slicks can be matched with the equivalent sections of the GCs of the source oil. If we can demonstrate the validity of this idea, software will be developed which could provide a powerful means of identifying and prosecuting polluters. But first we need to satisfy ourselves that humans are capable of identifying matches in the patterns. We are therefore seeking volunteers to participate in a visual pattern matching exercise.

The following pages contain a series of line graphs. No information is provided on the actual *meaning* of each graph - no titles, axes or labels are given; instead, we want you to focus only on the shape or *pattern* of each graph.

Each page contains three graphs (labelled 1 - 3) which we want you to match with one of the other six possible matching patterns (labelled A - F) according to which you think is the most similar. Your choices should be recorded in the table provided on each page. Additionally, you should indicate the degree of certainty you have about your chosen match (on a scale of 0 - 10, with 0 being completely uncertain and 10 being completely certain) - that is, a measure of how good a match you think it is.

For example, if you think:

- Graph 1 is best matched by pattern D, and this is a very close match
- Graph 2 is best matched by pattern D, although this is not a particularly close match
- Graph 3 is best matched by pattern B, and this is a fairly good match

then you could fill in the table as:

Graph	Matching Pattern	Degree of Certainty
1	**D**	10
2	**D**	4
3	**B**	7

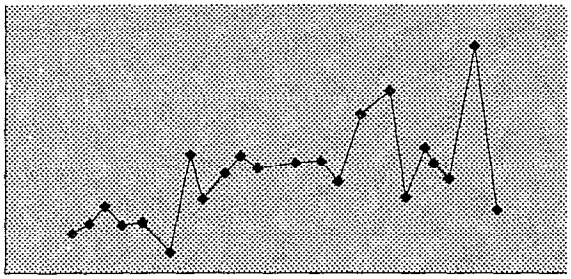
Note that each graph could match *any* of the given patterns A - F; ***there is no need to choose a different pattern for each graph.*** Please choose the BEST match in each case. (If you do not think *any* of the patterns are possible matches, indicate this with a question mark and/or degree of certainty zero.)

There is space at the end for you to write any comments (e.g. what difficulties you encountered in completing the task, any strategies you used when assigning matches).

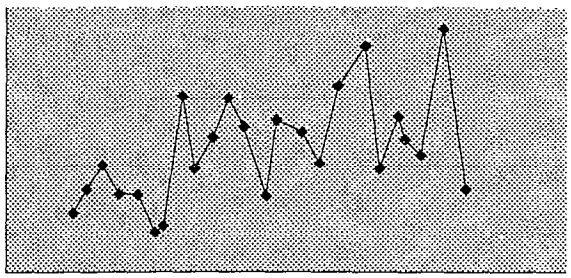
Thank-you for your time in completing this exercise.

Please return the completed forms to: Mark O'Connor or Bill Walley, as soon as possible but not later than Friday 18th December 1998. Address: School of Computing, Staffordshire University, Beaconside, Stafford, ST18 0DG. Tel: 01785 353510.

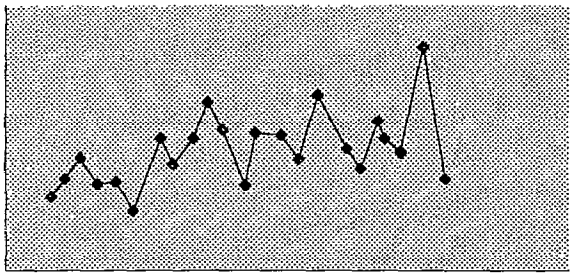
Graph 1



Graph 2

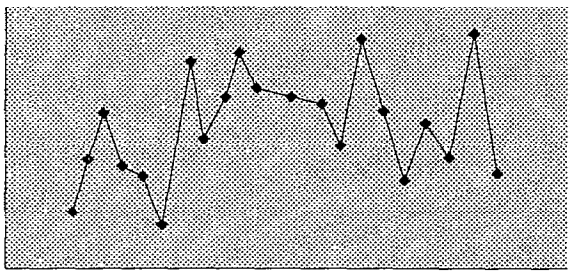


Graph 3

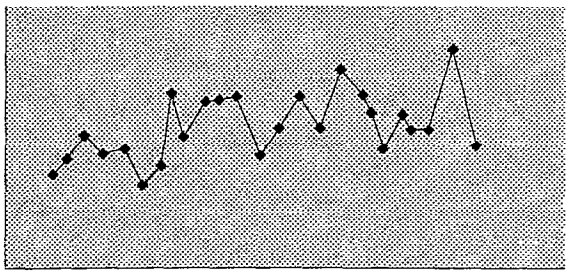


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

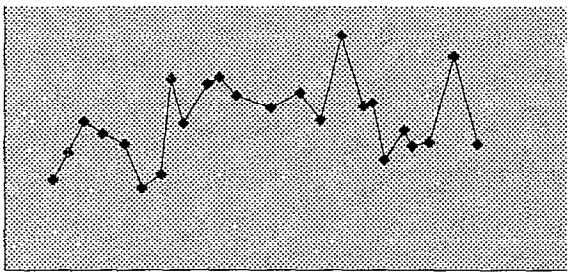
Pattern A



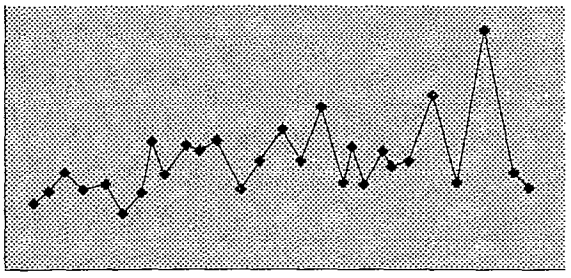
Pattern B



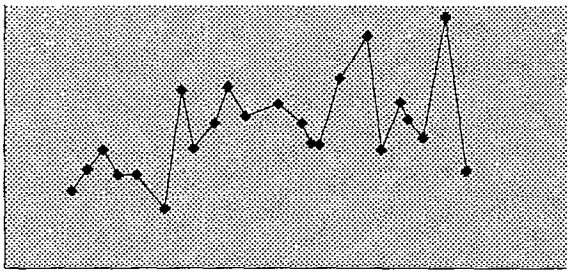
Pattern C



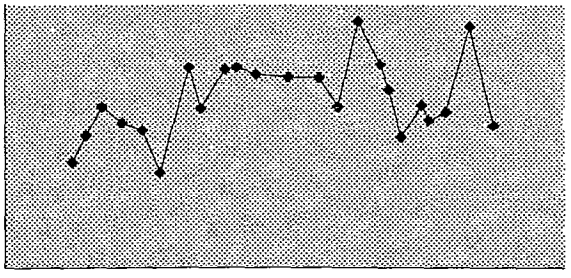
Pattern D



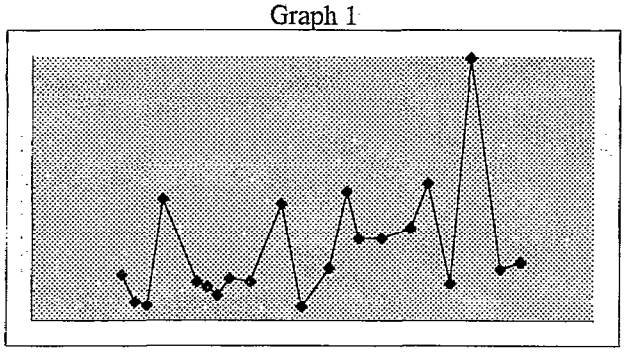
Pattern E



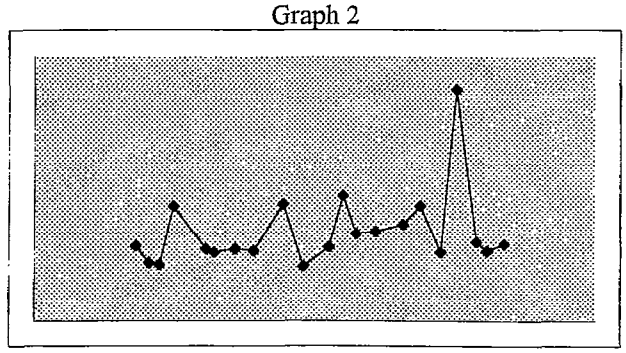
Pattern F



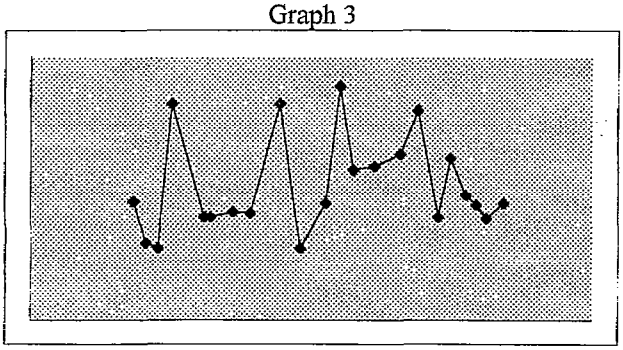
Graph 1



Graph 2

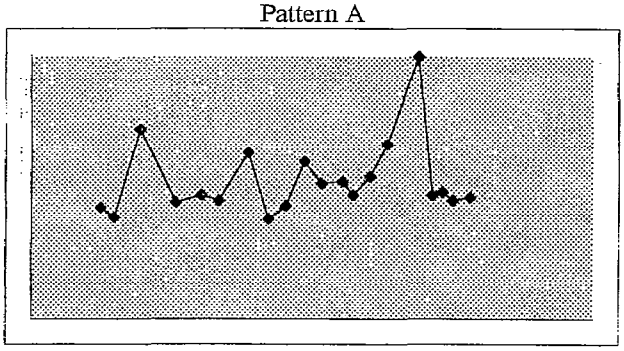


Graph 3

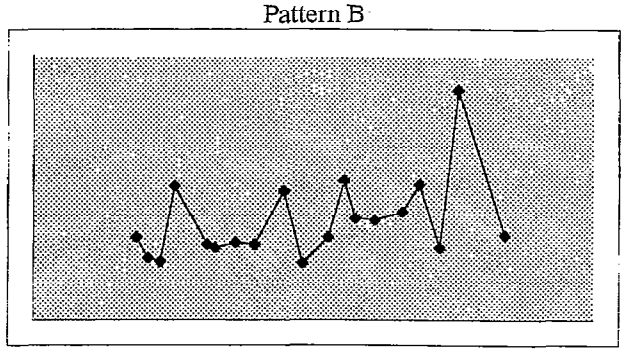


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

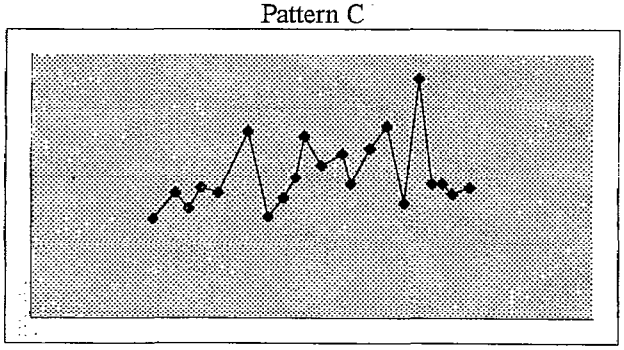
Pattern A



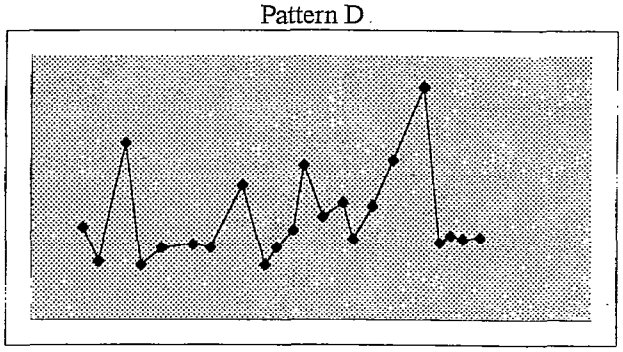
Pattern B



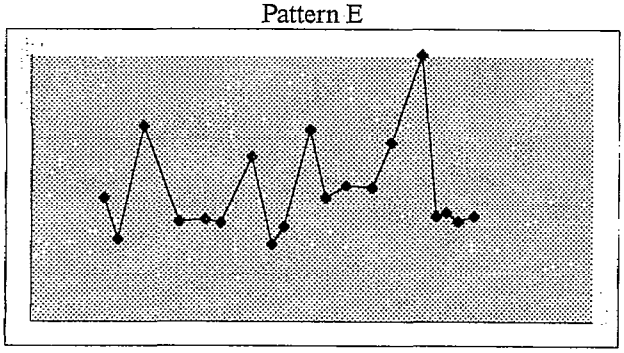
Pattern C



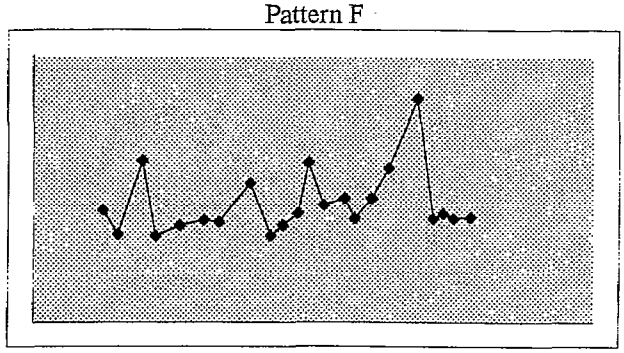
Pattern D



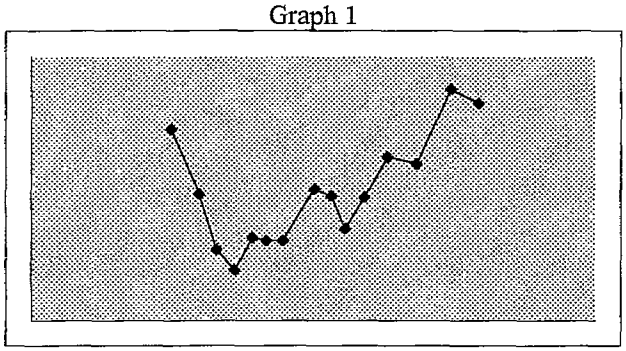
Pattern E



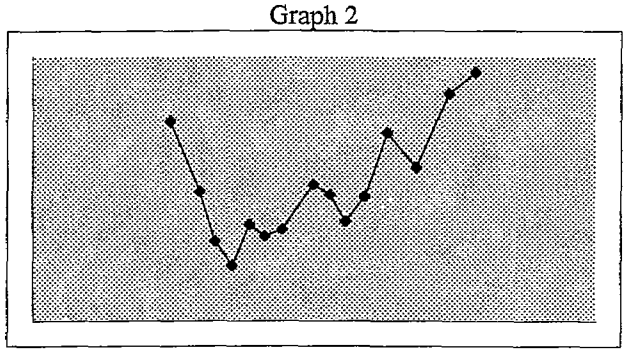
Pattern F



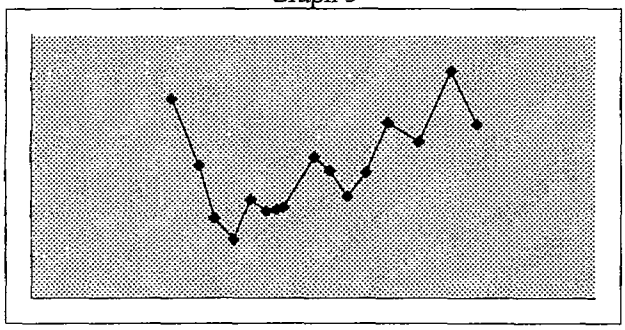
Graph 1



Graph 2

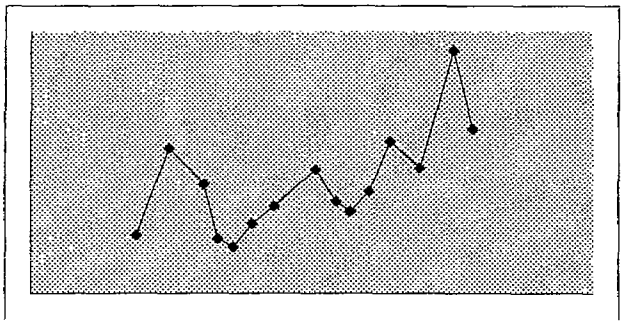


Graph 3

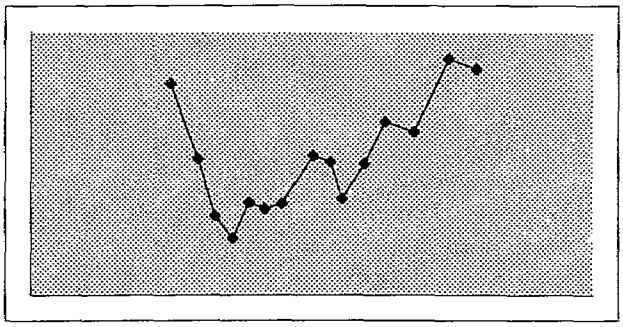


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

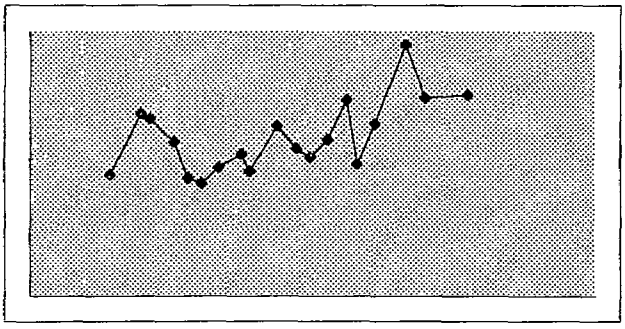
Pattern A



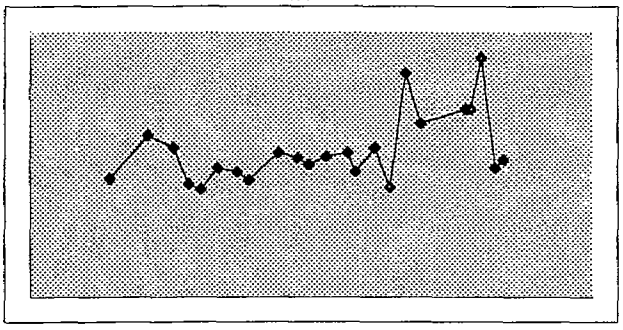
Pattern B



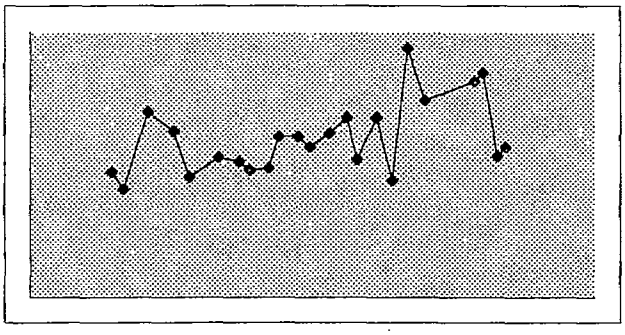
Pattern C



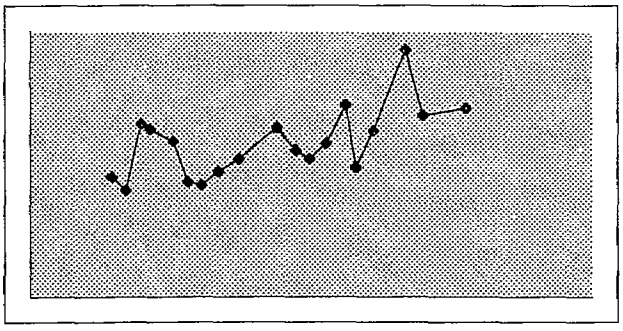
Pattern D



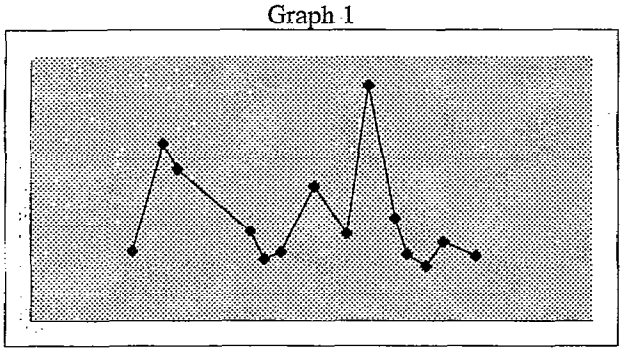
Pattern E



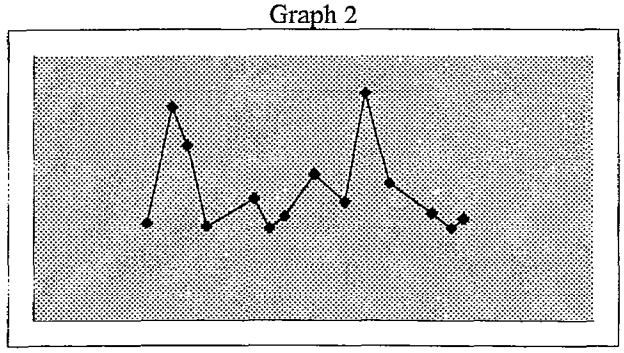
Pattern F



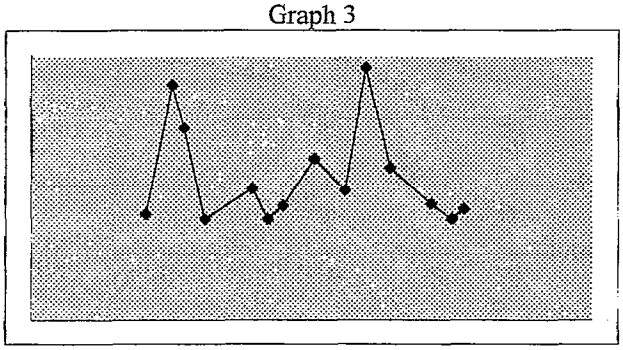
Graph 1



Graph 2

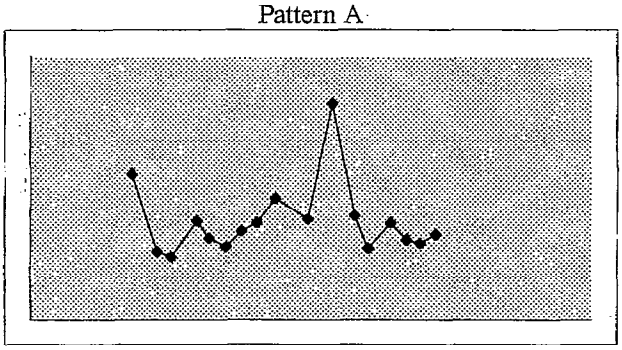


Graph 3

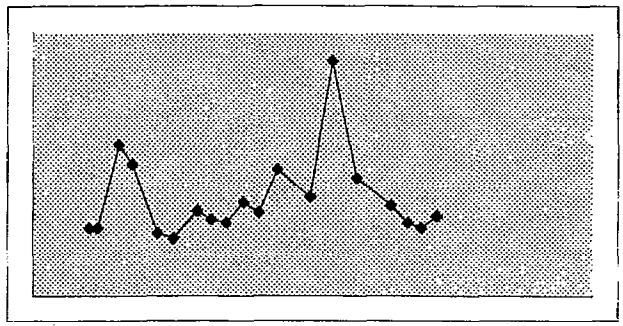


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

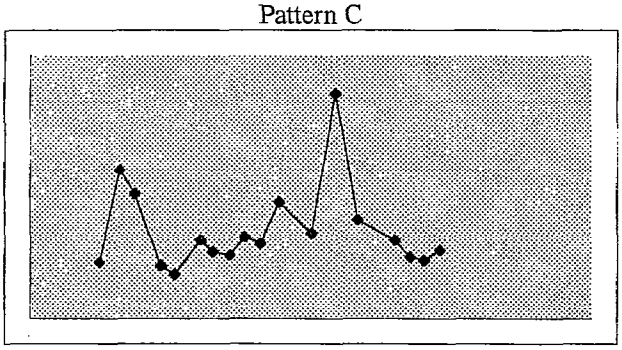
Pattern A



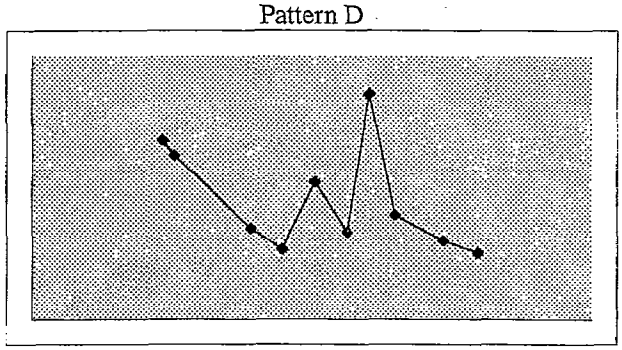
Pattern B



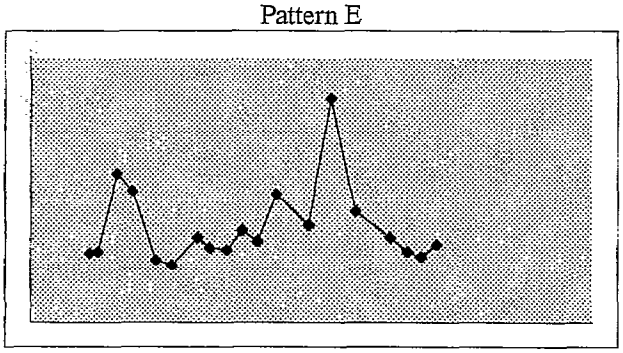
Pattern C



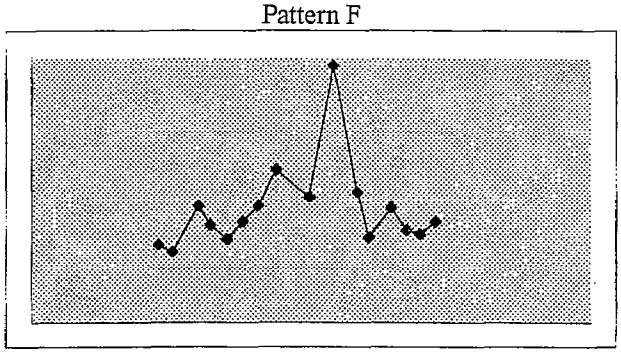
Pattern D



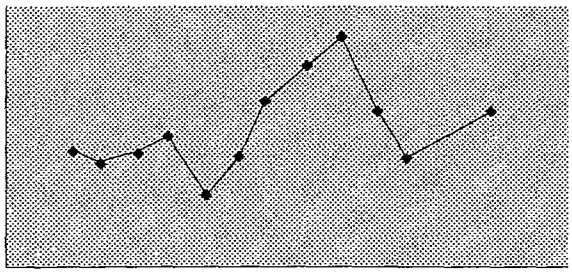
Pattern E



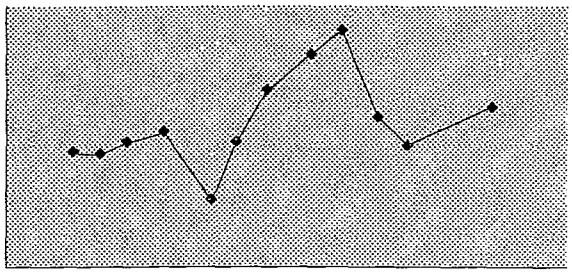
Pattern F



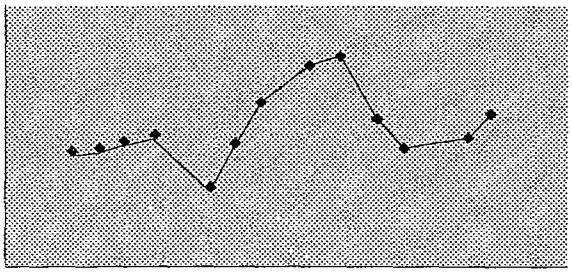
Graph 1



Graph 2

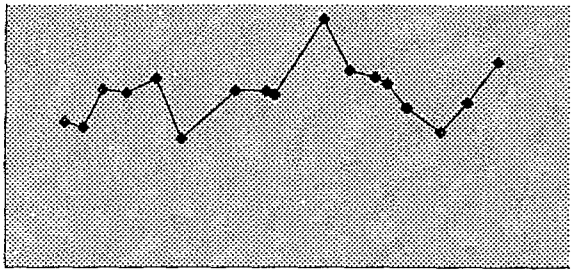


Graph 3

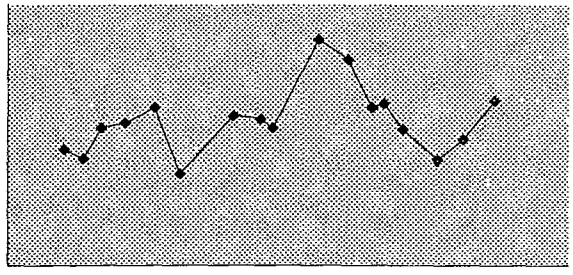


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

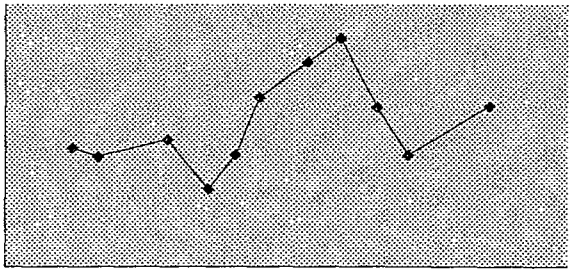
Pattern A



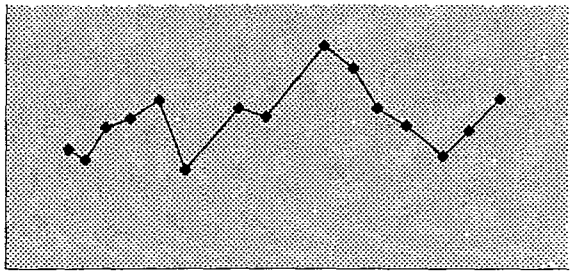
Pattern B



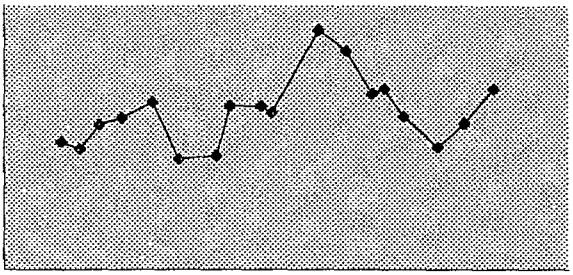
Pattern C



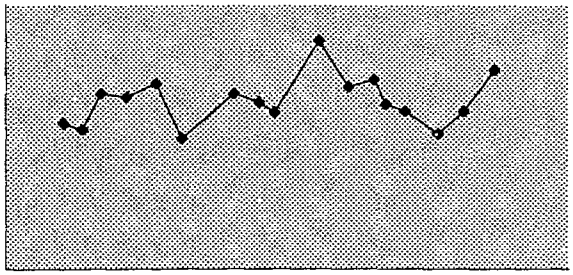
Pattern D



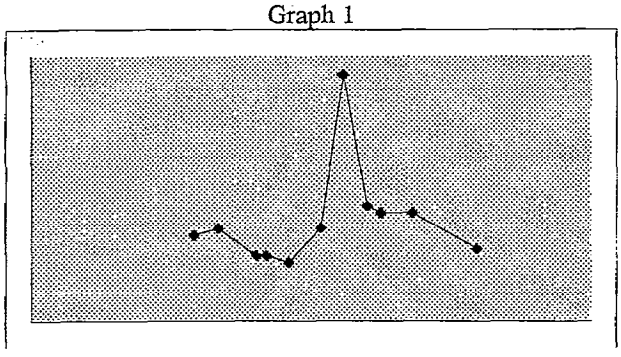
Pattern E



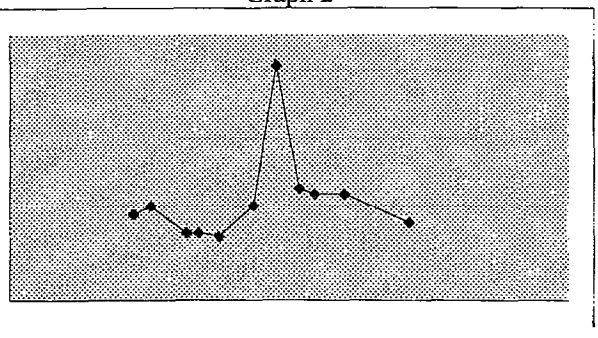
Pattern F



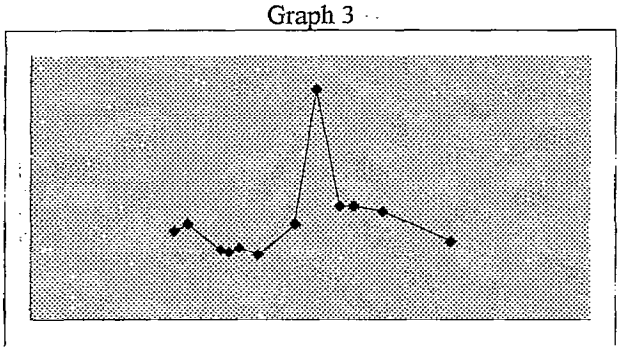
Graph 1



Graph 2

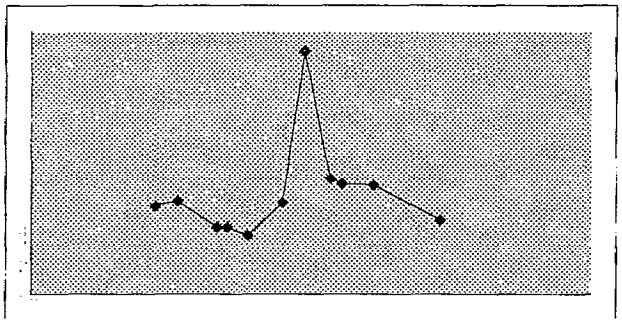


Graph 3

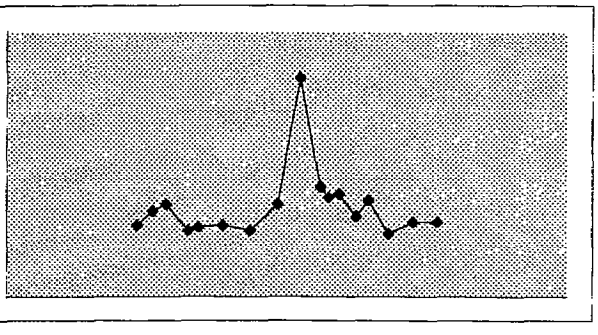


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

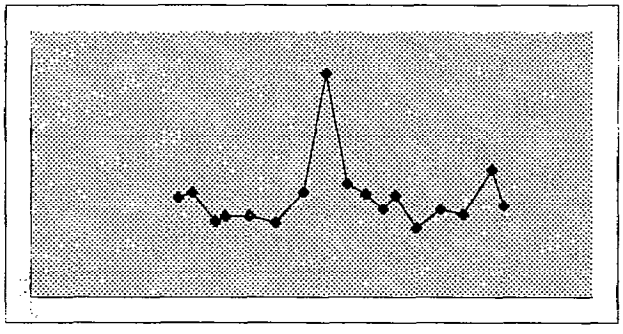
Pattern A



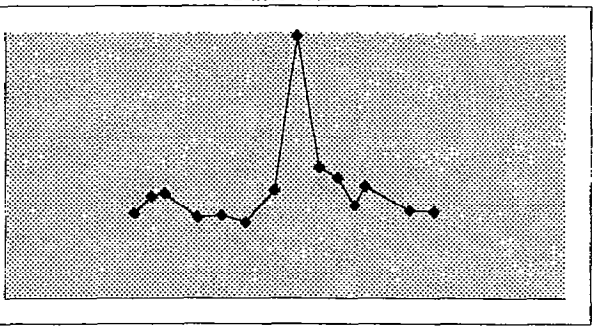
Pattern B



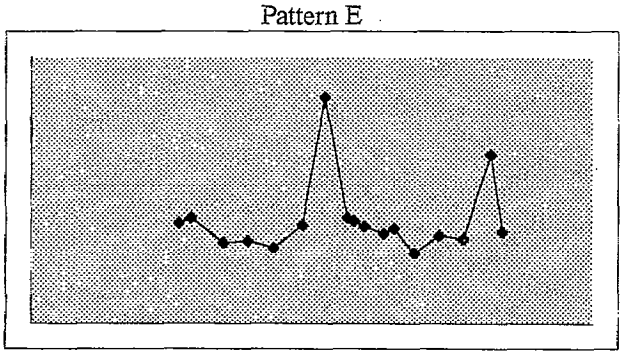
Pattern C



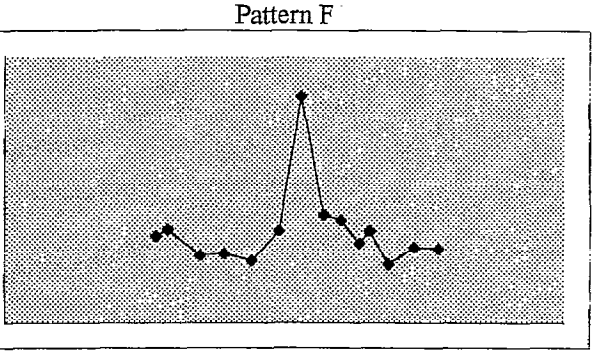
Pattern D



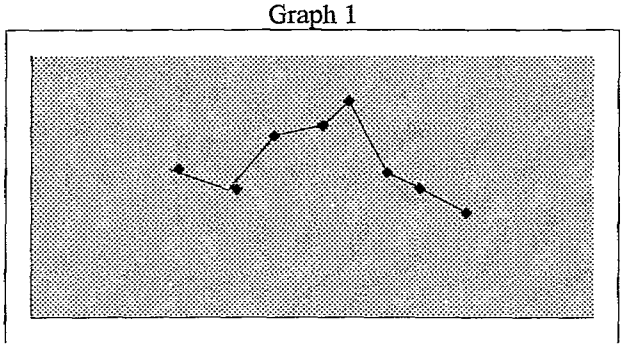
Pattern E



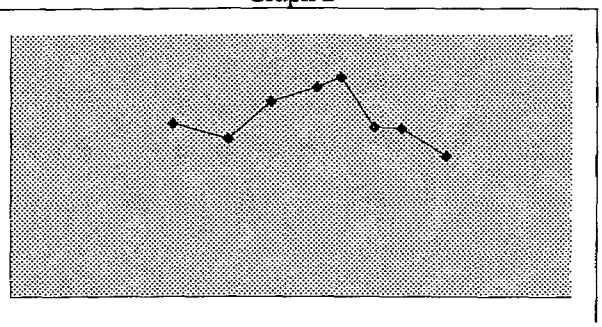
Pattern F



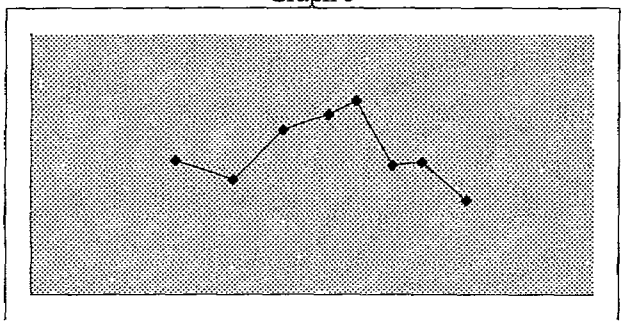
Graph 1



Graph 2

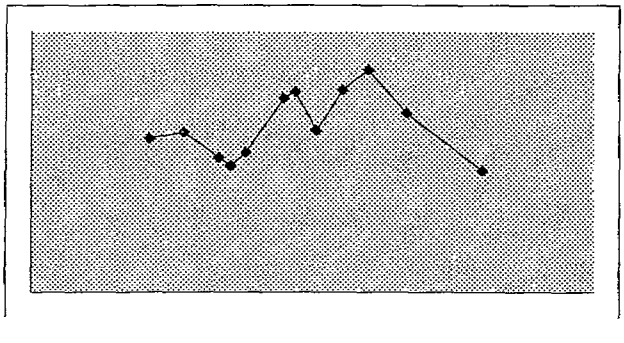


Graph 3

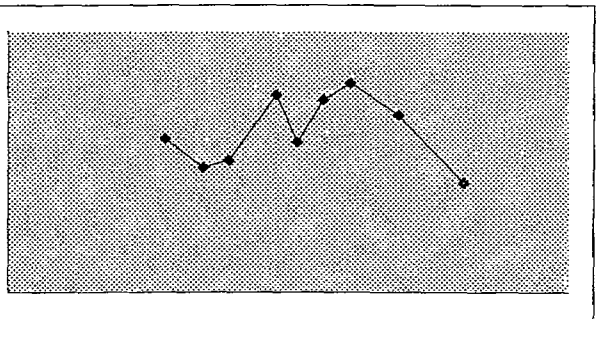


Graph	Matching Pattern	Degree of Certainty
1		
2		
3		

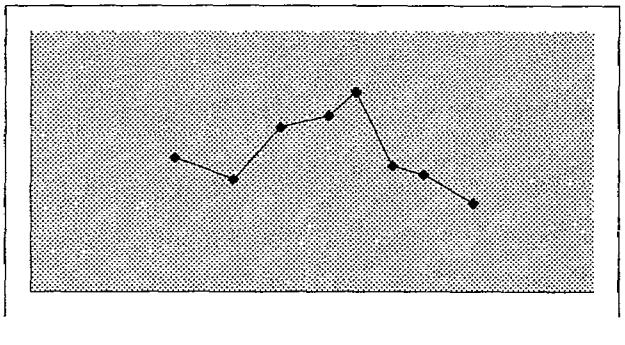
Pattern A



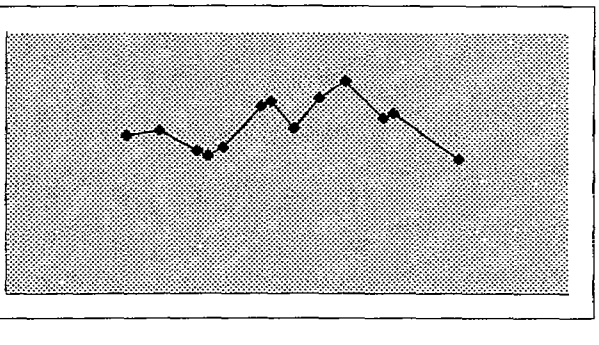
Pattern B



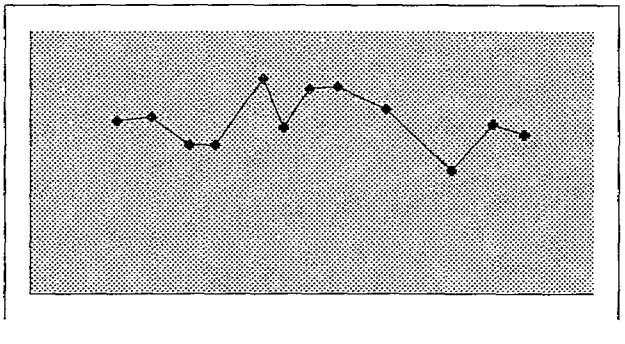
Pattern C



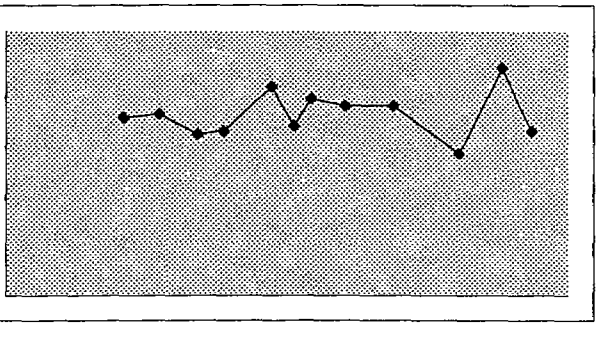
Pattern D



Pattern E



Pattern F



APPENDIX C

The Source Identification

Feasibility Study

-

Detailed Results

Notes on Table C1

The following table summarises the responses obtained from the pattern recognition exercise described in Section 5.2.

The first column contains the chromatogram section number (1 - 7), and also the code (A - F) for the correct match for that section (i.e. the letter representing the true source).

The second column contains the sample number (1 = water sample, 2 = surface sample, 3 = soil sample).

For each respondent (numbered 1 - 20 in the following columns), the letter representing the given match for the specified section and sample is listed; immediately below is the corresponding 'degree of certainty' assigned to this match by the respondent.

All 'incorrect' matches (including cases where no match was specified) are shaded.

The final four columns provide a summary of the results for each of the samples:

% corr. : Percentage of responses matching the true source sample.
Avg. cert. : Average certainty score allocated when matching a sample.
(corr resp) : Average certainty score allocated when a sample was matched to the true source.
(inc. resp) : Average certainty score allocated when a sample was either matched to a source other than the true source, or no match was specified.

Similarly, the final four rows provide a summary of the results for each of the respondents.

Average values for each of the final four rows and columns are also given.

Table C1. Full results from pattern matching exercise: 20 respondents' choices and 'certainty' scores.

		Respondent Number																							
Section no.	Sample no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	% corr.	Avg. cert.	(corr resp)	(inc resp)
1 (E)	1	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	100			
		6	9	8	8	5	8	6	10	7	8	9	9	9	6	10	7	10	8	9	8		8.00	8.00	-
	2	E	C	E	E	E	E	E	E	E	E	E	E	A	E	X	E	E	E	E	E	85			
		5	4	9	7	6	5	3	4	6	7	9	9	7	5	0	6	7	7	8	7		6.05	6.47	3.67
	3	E	B	E	E	E	E	B	B	B	A	E	E	B	E	X	B	E	C	E	E	55			
		2	6	7	5	5	6	4	2	6	8	8	7	7	3	0	4	7	7	8	5		5.35	5.73	4.89
2 (B)	1	B	E	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	95			
		7	7	8	6	6	7	4	8	6	9	9	9	8	6	5	7	8	9	7	8		7.20	7.21	7.00
	2	B	B	B	B	B	B	B	B	F	B	B	B	C	B	B	B	B	B	B	B	90			
		6	8	9	7	7	8	7	8	6	10	8	9	7	9	5	7	7	9	8	8		7.65	7.78	6.50
	3	B	D	E	B	B	E	B	E	E	B	E	A	D	E	X	X	E	E	E	X	25			
		2	6	5	3	5	6	3	4	5	8	5	5	6	2	0	0	8	6	8	0		4.35	4.20	4.40
3 (B)	1	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	100			
		10	10	10	9	10	9	8	8	8	10	10	10	10	10	5	9	10	10	9	8		9.15	9.15	-
	2	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	100			
		9	8	8	6	8	8	5	8	6	10	9	9	9	8	2	8	10	9	8	7		7.75	7.75	-
	3	B	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	B	B	90			
		9	4	9	5	6	8	6	6	6	9	8	9	10	7	2	3	8	8	9	7		6.95	7.06	6.00
4 (D)	1	D	C	C	C	E	C	X	D	E	C	C	D	F	E	B	B	C	E	C	D	20			
		3	5	7	6	4	6	0	4	5	6	6	6	7	2	1	4	4	7	8	6		4.85	4.75	4.88
	2	C	C	E	B	C	C	C	X	B	C	C	C	B	C	X	C	B	C	C	X	0			
		1	6	7	6	7	8	6	0	7	7	6	6	7	4	0	7	4	8	8	0		5.25	-	5.25
	3	B	D	C	C	C	C	C	X	C	C	C	C	C	C	X	C	E	E	C	X	5			
		2	4	8	5	5	7	5	0	7	7	6	6	7	2	0	7	4	6	8	0		4.80	4.00	4.84
5 (C)	1	C	C	C	C	C	C	C	C	E	C	C	C	C	C	C	C	C	C	C	C	95			
		10	9	9	9	9	10	7	10	5	10	9	10	9	9	5	7	5	10	6	9		8.35	8.53	5.00
	2	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	D	C	95			
		10	9	9	8	6	9	5	10	7	9	9	9	9	8	5	8	4	8	5	9		7.80	7.95	5.00
	3	C	D	C	C	C	C	X	C	B	C	C	C	C	C	C	C	D	C	D	C	75			
		8	6	8	6	5	8	0	8	5	8	9	9	8	7	5	3	2	8	5	9		6.35	7.27	3.60
6 (A)	1	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	100			
		10	10	10	10	10	9	9	10	8	10	10	10	9	10	10	10	8	9	10	9		9.55	9.55	-
	2	A	A	A	A	A	A	A	A	A	A	A	A	F	A	A	A	A	A	A	A	95			
		10	9	9	9	9	9	9	10	8	10	10	10	7	10	10	9	8	10	9	9		9.20	9.32	7.00
	3	A	A	A	A	A	A	A	A	A	A	A	A	D	A	A	A	A	A	A	A	95			
		9	9	9	8	6	9	8	10	8	9	9	8	7	10	8	8	8	9	7	9		8.40	8.47	7.00
7 (C)	1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	100			
		9	10	9	9	9	10	8	10	7	10	10	10	10	8	8	8	8	8	9	10		9.00	9.00	-
	2	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	100			
		7	9	9	6	6	8	5	8	5	9	9	9	10	7	10	6	8	7	9	9		7.80	7.80	-
	3	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	100			
		9	10	9	7	7	8	6	8	6	9	9	9	10	7	9	6	8	6	9	9		8.05	8.05	-
% correct		90.5	61.9	81.0	85.7	85.7	81.0	76.2	81.0	61.9	81.0	81.0	85.7	57.1	81.0	71.4	76.2	71.4	76.2	71.4	85.7	77.1		7.4	5.4
Avg certainty		6.9	7.5	8.4	6.9	6.7	7.9	5.4	7.0	6.4	8.7	8.4	8.5	8.2	6.7	4.8	6.4	7.0	8.0	8.0	7.0		7.2		
...(corr resp)		7.4	8.8	8.8	7.1	6.9	8.2	6.2	8.2	6.8	9.1	9.1	8.9	9.3	7.6	6.6	7.0	7.7	8.4	8.3	8.1	7.9			
...(incorr resp)		1.5	5.5	6.8	5.7	5.3	6.8	3.0	1.5	5.8	7.0	5.8	5.7	6.9	2.5	0.2	4.4	5.0	6.8	7.0	0.0	4.6			