Numerical methods for comparing fresh and weathered oils by their FTIR spectra

Jianfeng Li,*[a] D. Brynn Hibbert[a] and Stephen Fuller[b]

Four comparison statistics ('similarity indices') for the identification of the source of a petroleum oil spill based on the ASTM standard test method D3414 were investigated. Namely, (1) first difference correlation coefficient squared and (2) correlation coefficient squared, (3) first difference Euclidean cosine squared and (4) Euclidean cosine squared. For numerical comparison, an FTIR spectrum is divided into three regions, described as: fingerprint (900–700 cm^{-1}), generic (1350–900 cm^{-1}) and supplementary (1770–1685 cm^{-1}), which are the same as the three major regions recommended by the ASTM standard. For fresh oil samples, each similarity index was able to distinguish between replicate independent spectra of the same sample and between different samples. In general, the two first difference-based indices worked better than their parent indices. To provide samples to reveal relationships between weathered and fresh oils, a simple artificial weathering procedure was carried out. Euclidean cosine and correlation coefficients both worked well to maintain identification of a match in the fingerprint region and the two first difference indices were better in the generic region. Receiver operating characteristic curves (true positive rate *versus* false positive rate) for decisions on matching using the fingerprint region showed two samples could be matched when the difference in weathering time was up to 7 days. Beyond this time the true positive rate falls and samples cannot be reliably matched. However, artificial weathering of a fresh source sample can aid the matching of a weathered sample to its real source from a pool of very similar candidates.

Introduction

Faced with an environmental oil spill, environment protection authorities commonly use one or all of three analytical techniques to compare the spill with candidate sources: gas chromatography/mass spectrometry (GC-MS), Fourier transform infrared (FTIR) and fluorescence spectra. Of these, GC-MS on biomarkers is most used to claim definitive matches, but FTIR and fluorescence spectra represent quicker and cheaper ways of screening samples, and are often used as a precursor to GC-MS analysis. Petroleum-based oil and its refined products are mixtures of hydrocarbon compounds. Light refined products, such as gasoline and kerosene, evaporate quickly and leave little residue in the environment, but heavier refined oil products, like diesel fuel, are more persistent and may have long-term effects on human and wildlife in the polluted area.[1,2] Oils are sufficiently complex mixtures, that bear signatures of their origin and refining, so it is expected that two fresh samples of an oil, with proper sampling and analysis, should be able to be matched if indeed they come from a common source. However, when a sample of oil is exposed to the environment, its composition changes with time due to weathering. Matching weathered samples to fresh

oils adds complexity because the analyst must determine which differences are due to weathering and which are due to genuine differences between the samples. Statistics[3,4] and pattern recognition[5,6] have been used for comparing IR, fluorescence spectra or gas chromatograms, and recently we have shown the use of different matching statistics for fluorescence spectra.[7] In particular cases, systematic samples can be collected and analysed by multivariate analysis techniques.[8,9] However, for routine bench analysis, this approach is not possible because of the lack of suitable training data sets.

ASTM International's active standard—ASTM D3414,[10] which describes a method for comparing the IR spectrum of a spilled oil with spectra of suspect sources by a peak-to-peak comparisons, has been used in the laboratory of the Department of Environment and Conservation of NSW (DEC, NSW) for years. It has been proved to be authoritative and useful. There are prerequisites for using this standard technique, including purity of sample, comparability of spectral baseline and similar thicknesses of samples in the sample cell. The major information region of a spectrum of an oil is between 2000 to 600 cm^{-1}. Within this range the "oil fingerprint" region between 900 and 700 cm^{-1} includes detailed characteristics while the region 1350 to 900 cm^{-1} has stronger link to the general effects of weathering. Weathered oils are usually oxidized, leading to the rise of a carbonyl peak between 1770 to 1685 cm^{-1}, a feature which the spectra of many fresh oils do not have. The ASTM interpretation procedure starts from prerequisite checks of thickness of samples, then determines effects of weathering, and finally,

[a]*School of Chemistry, University of New South Wales, Sydney, NSW, 2052, Australia. E-mail: jianfeng.li@unsw.edu.au; Fax: +61-293856141; Tel: 61-293854821*
[b]*Environmental Forensic and Analytical Science Section, Department of Environment and Conservation, PO Box 29, Lidcombe, NSW, 1825, Australia*

carefully compares spectral similarities in the fingerprint region. For visually comparing two spectra, the standard suggests superimposing two spectra using a light-box or recording two spectra on the same chart. The standard lists a number of peaks which should be checked for different types of oils to decide if differences arise from weathering. Consideration of weathering effects on spectra defines criteria for deciding there is a "probable match" when there is not perfect agreement between the spectra (Table 2 in ref. 10). The ultimate degree of match of two spectra depends on the match of overall shapes and detailed differences between two spectra, a procedure requiring a great deal of judgment by the analyst. In the authors' opinion, a junior analyst would find it difficult to follow the standard and finish a comparison between two oils. Another issue is that the purity of the environmental sample is not always guaranteed. The spill sample almost always has interferences from the matrix of the spill site (more in the case of land spills, and less in waterborne pollution) and carries the changes triggered by exposure to the environment. The judgment of what is a reasonable difference between spectra that arises from the effect of interferences is subjective and in complex cases could also challenge the experience of the analyst.

Since the D3414 provides a sound and proven guidance for source identification of oil spill based on FTIR spectroscopy, we propose numerical procedures that are based on the methodology of the Standard as alternative means of the current method. Because correlation coefficient has been used widely for various interests, in this paper we report the usability of it and related numerical similarity indexes for matching oil samples. Comparing different fresh samples and fresh with weathered samples were used as two scenarios and practical aspects of using these numerical methods are discussed.

Experimental

Oil samples

All oil samples were provided by the Department of Environment and Conservation of New South Wales and ASTM protocol[10] for sampling has been followed during all sample preparations. These were typical diesels used as automotive fuel and a crude oil and no further information was available. Three typical diesel oils, used as representative similar oils from different sources, are coded I, III and IV and a crude oil is coded sample II in the text.

Weathering experiment

Weathering[11,12] of oil in natural environment includes a variety of processes, including evaporation,[13] dissolution,[14] and photochemical oxidation.[15] Different biological environments on land and water also give rise to different outcomes of weathering.[16] It is almost impossible to reproduce the weathering conditions of any given spill. The weathering protocol used in our investigation was simple, mainly involving photooxidation and evaporation, and was only intended to provide reasonably representative samples to investigate the numerical procedures developed here. The weathering procedure for the oil samples I and IV was to form a 2 to 5 mm thick slick of the oil over water in an open beaker. The beakers were placed in an unprotected area on the roof of a building for 2, 7 or 15 days from 21st November to 4th December, 2002. During this time in Sydney, Australia the weather was dry and hot with three days having a maximum temperature over 35 °C. Results from these samples are coded with the number of the oil for a fresh sample, suffixed with w2d (2 days), w7d (7 days) and w15d (15 days) for increasingly weathered samples.

FTIR spectroscopy

Infrared spectra of each fresh oil sample were collected on a Fourier transform infrared spectrophotometer (Excalibur FTS 3000, Bio-Rad Laboratories, Hercules, CA, USA). The oil samples were analysed under the same conditions using the same KBr cell which was cleaned between samples, and the spectra were recorded from 4000 to 650 cm^{-1}. 32 scans at a resolution of 4 cm^{-1} were collected and averaged for the background and for each sample. Each spectrum had 1738 data points. A 0.05 mm spacer in the cell ensured consistent thickness of the oil sample therefore no correction for thickness was needed.

The same configuration was used to analyse samples I and IV and their weathered samples at the University of New South Wales using a Nicolet Avatar System 370 (Thermo-Nicolet, Madison, WI, USA).

Software

All calculations were performed with MATLAB, version 2006a for Windows (The MathWorks, Natick, MA, USA), using scripts written by the authors.

Spectra pre-processing

The spectra of the oils were all very clean and the peaks were distinct. For the results given here, baselines were approximated by linear functions which were fitted to three non-peak regions, 646 cm^{-1} to 669 cm^{-1}, 1851 cm^{-1} to 2333 cm^{-1} and 3394 cm^{-1} to 3973 cm^{-1}. These functions were then used to correct for baseline in the spectral regions of interest. Corrected spectra were transformed to transmittance for later comparison. The raw data were also analysed by the methods given here, and no significant difference between the results (pre-processed and not pre-processed) was found.

Comparison procedure

Based on the ASTM standard for comparing two FTIR spectra, the method divides a spectrum into three regions: a fingerprint region between 900 and 700 cm^{-1}, a generic region between 1350 and 900 cm^{-1} and a supplementary weathering region between 1770 and 1685 cm^{-1}. Using these divisions does not only follow the ASTM standard but also ensures that the correlation coefficient related methods work properly. In our experience, inclusion of non-informative segments or mixing different informative regions decreases the performance of numerical methods.

The spectral fragments in the first two regions were compared using each of the point-to-point comparison methods

described below. Because there is only one major peak in the supplementary weathering region, arising from oxidation of hydrocarbons giving a carbonyl group, only the absolute difference of the average transmittance in the region was calculated. While the fingerprint region is the key region for a comparison, the generic region can provide supportive information. For the comparison of two fresh oils without weathering effects, only the first two regions are needed by the ASTM specification. In all comparison regions, transmittance is used in preference to absorbance, because although the latter is proportional to amount concentration, transmittance gives greater weight to minor components that are of importance in discriminating between oils.

The four comparison methods used in this study calculate the following similarity indices: (1) first difference correlation coefficient squared (1stCor); (2) correlation coefficient squared (Cor); (3) first difference Euclidean cosine squared (1stCos); and (4) Euclidean cosine squared (Cos). The defining equation of each index is given in Table 1. All similarity indices used here have values in the interval [0,1], from 0 (no match) to 1 (perfect match).

In first difference-based methods, the differences between successive transmittances are calculated for each spectrum, and comparison is then made using the equations given in Table 1. Numerical estimates of the first derivatives of the spectra are thus compared. The use of first difference-based similarity indices normally improve baseline artefacts, *i.e.* differentiation of conjoint transmittances of a spectrum removes a constant baseline, renders a linear slope to a constant and, in general, reduces the dimension of the baseline by one. Such processing also sharpens features, allowing differences between similar spectra to be distinguished.

The sample set analysed included four fresh (samples I to IV) oils, and oils I and IV each also had three weathered samples (coded with suffixes w2d, w7d and w15d) in different weathering degrees. All these ten oils were each independently sampled at least 10 times to acquire a pool of spectra of the each oil in a particular state. For comparing a pair of oil samples, every spectrum of the replicas of one sample were compared to every spectrum of the replicas of another sample. Each comparison of a pair of samples created at least 45 similarity indices which gave relatively enough data for creating a smooth distribution of the similarity indices of a comparison. The median of the indices was used when a single representative number was needed, otherwise they were used individually.

ROC curves

A receiver operating characteristic (ROC) curve[17,18] is a plot of a comparison method's true positive rate (TPR) or sensitivity *versus* its false positive rate (FPR) for particular decision thresholds. The TPR is the probability that samples come from the same source are correctly identified as a match. The FPR is the probability of false classification, *i.e.* that non-matched samples will be incorrectly classified as a match. Each point on the graph is calculated at a different decision threshold (value of the similarity index). The slope from the origin to a point on the curve is the likelihood ratio of the probability of a match to the probability of a no match for a pair of spectra with a similarity index. As the TPR increases, the FPR also tends to increase, and from the plot it is possible to identify the thresholds that give the highest possible TPR, the lowest FPR or the minimum total error ($1 - $ TPR + FPR). Once a ROC curve is available, an analyst can use any point on the curve to decide a decision threshold depending on the relative risk between false positive and false negatives.

To obtain a ROC curve for matching weathered samples, a test set of samples is analysed for which the origin and subsequent weathering history of the material is known. The similarity index is calculated for every pair of spectra. A 'matched' pair of spectra is defined as two spectra from a given oil irrespective of weathering history. 'Non-matched' pairs of spectra were chosen from one spectrum of the fresh oil that provided the matched pairs, and a spectrum of an oil of different origin irrespective of weathering history. In a spreadsheet, the values of the similarity indices from matched and non-matched pairs are ordered while maintaining, in an adjacent column, information about the true state of the match ('matched' or 'non-matched'). Starting with the smallest value of the index the FPR is the count of the remaining 'non-matched' values divided by the total of non-matches, and similarly the TPR is the count of the remaining matched values divided by the total matched spectra. For less than the smallest value of the similarity index observed in the set, TPR = FPR = 1, *i.e.* the decision value is such that all spectra are called a match. As the index of the decision point increases, hopefully, non-matched pairs are rejected, lowering the FPR, and ideally all of these are passed over before any true positive is incorrectly rejected thus maintaining a TPR of 1 as the FPR falls to zero. In practice, there is often an overlap region in which the shape of the ROC curve can be used to optimize the decision point. While the ROC curve lies above the 45° line, the method does give information that is better than a random choice as to whether a pair matches or not.

Table 1 Definitions of similarity indices. A_i are the vectors of the instrument responses for sample i ($i = 1, 2$), with n elements $A_{i,j}$, $j = 1,\ldots, n$, which can be absorbance or transmittance. ΔA_i is the first order difference for sample i which is defined as: $\Delta A_{i,j} = (A_{i,j+1} - A_{i,j})$, $j = 1,\ldots, n - 1$. cov(A_1,A_2) and cov($\Delta A_1,\Delta A_2$) are the covariance of the vectors of the raw and the first order differenced spectra to be matched, respectively, and σ is the standard deviation of a spectrum

First difference correlation coefficient squared	$\left(\dfrac{\mathrm{cov}(\Delta A_1,\Delta A_2)}{\sigma_{\Delta A_1}\sigma_{\Delta A_2}}\right)^2$
Correlation coefficient squared	$\left(\dfrac{\mathrm{cov}(A_1,A_2)}{\sigma_{A_1}\sigma_{A_2}}\right)^2$
First difference Euclidean cosine squared	$\left(\dfrac{\sum_{j}^{n-1}\left(\Delta A_{1,j}\Delta A_{2,j}\right)}{\sqrt{\sum_{j}^{n-1}\Delta A_{1,j}^2 \sum_{j}^{n-1}\Delta A_{2,j}^2}}\right)^2$
Euclidean cosine squared	$\left(\dfrac{\sum_{j}^{n}\left(A_{1,j}A_{2,j}\right)}{\sqrt{\sum_{j}^{n}A_{1,j}^2 \sum_{j}^{n}A_{2,j}^2}}\right)^2$

The analysis here is performed on the data itself which generates a ROC curve of discrete points. It is also possible to determine the distributions (probability density functions) of matched and non-matched values and then generate a ROC curve from the continuous analytical functions. For some applications the area under the ROC curve (AUC) is a good indicator of the quality of the method.[19] Here, however, it is necessary to scrutinise the entire curve.
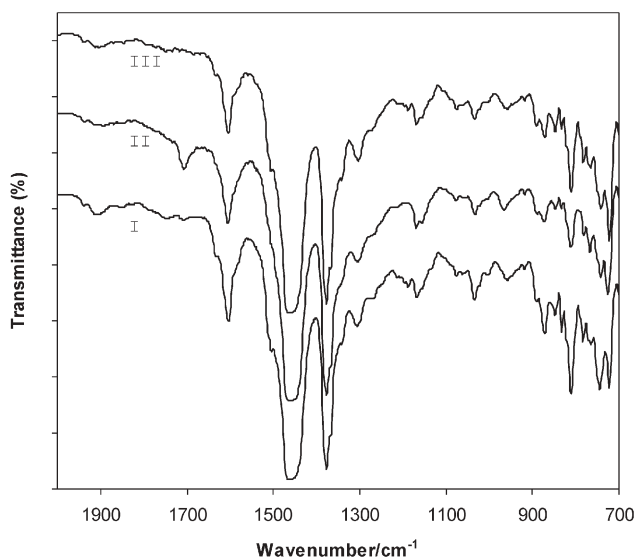
Results

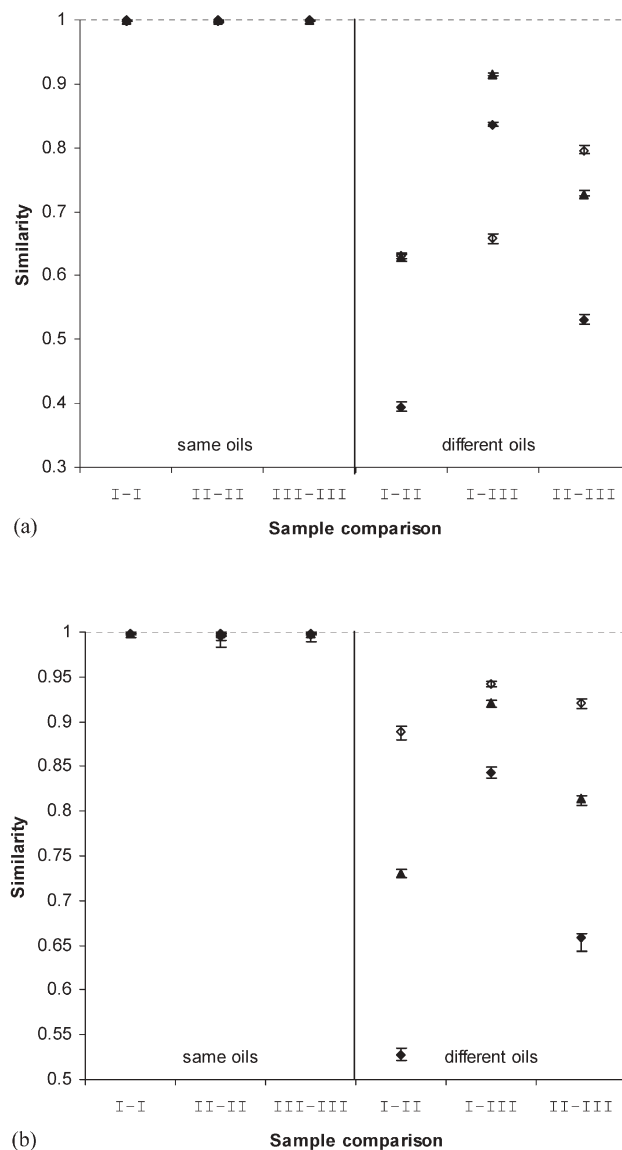### Scenario one: comparison of different fresh samples

Two diesel samples I and III, and a crude oil sample II were used to test the ability of the methods to distinguish between different fresh samples. Ten replicate independent test portions of each oil were analysed in a KBr cell with a spacer. The average FTIR spectrum of each sample is shown in Fig. 1.

Visual inspection reveals some similarities and differences of the spectra, but it is not entirely obvious how the relations among them should be classified. A numerical method should give similarity indices near 1 for replicates of the same oil, and lesser values for comparisons between different oils. It may also be expected that the two diesels would be more similar to one another than a diesel and the crude oil. The two regions (fingerprint and generic) were compared separately and the results are shown in Fig. 2(a) and (b) respectively. The narrow range of Euclidean cosine squared, from 0.988 to 1 in the fingerprint region, and 0.997 to 1 in the generic region, for both similar and different samples makes it not suitable for the practical use of the method for comparing fresh samples, and these results have not been graphed in Fig. 2.

As might be expected, the results of the comparisons between the replicates of the same sample distribute in a very narrow range and have values close to 1 whilst the similarity indices of the comparisons between different samples range from 0.394 (median 1stCor for I–II) to 0.914 (median 1stCos

(a)

(b)

Fig. 2 Similarity indices between three oil samples (I, III, diesel; II, crude) calculated pair-wise among 10 replicate spectra of each oil: (a) in the fingerprint region (900–700 cm^{-1}) and (b) generic region (1350–900 cm^{-1}). In each diagram, the first three are the comparisons of the replicates of the same sample and the last three are the comparisons between different samples. The error bars and points represent the interquartile ranges about the medians. Filled diamonds: first difference correlation coefficient squared (1stCor), open diamonds: correlation coefficient squared (Cor), filled triangles: first difference Euclidean cosine squared (1stCos).

Fig. 1 Average FTIR spectra of diesel oil samples I, II and III from measurements on ten independent test portions. Spectra are overlaid with 30% transmittance offset for comparison.

for I–III) in the fingerprint region and 0.528 (median 1stCos for I–II) to 0.942 (median Cor for I–III) in the generic region. In the fingerprint region, for the first difference methods, and for all methods in the generic region, the results are as expected, with oils I and III (both diesels) being most similar. It is obvious that the spectra of all three samples are very similar in the fingerprint region. Correlation coefficients for the comparisons in the fingerprint region were in a narrower range, from 0.630 (median I–II) to 0.796 (median II–III). This method responds less to differences between spectra than its

first differentiated counterparts. Compared to the fingerprint region, the differences in generic regions are clearer and all methods can provide correct assignments. In both regions, 1stCor gave the biggest variance of the scores and correlation coefficient square gave the smallest. These results show the differences of discrimination ability between the methods. We have shown that similarity indices of a number of pair-wise spectral comparisons do not have a normal distribution[4], and therefore show in Fig. 2 medians and interquartile ranges rather than means and standard deviations. Overall, the first difference correlation coefficient squared gives the best results, and, in the example used, can unambiguously distinguish among same and different samples of diesels, and same and different kinds of oil.

## Scenario two: comparison of weathered samples

Weathering is a complex process with physical, chemical and biological effects. Weathering changes FTIR peaks more dramatically in some regions than in other regions. Eventually the whole spectrum could be very different to the spectrum of the fresh oil, but in general the rate of change of a sample undergoing weathering falls with time, as less material that can be weathered is available. Weathering provides a challenging test of the numerical comparison methods and it is expected to see the similarity indices change differently to weathering effects in different regions. Numerical matching methods should be sensitive to genuine differences and robust to weathering at the same time.

Samples of the two diesel oils I and IV, which have similar FTIR spectra (see Fig. 3), were subjected to the weathering procedure described above at the same time.

The averages of at least 10 spectra of increasingly weathered samples of oil IV are shown in Fig. 4. The growth of the peak at 1714.5 cm^{-1}, movements of the baseline and minor changes in the fingerprint region may be seen that are consistent with the effects of weathering. Some useful discussion on weathering effects on oils can be found in ASTM D3414.[10]
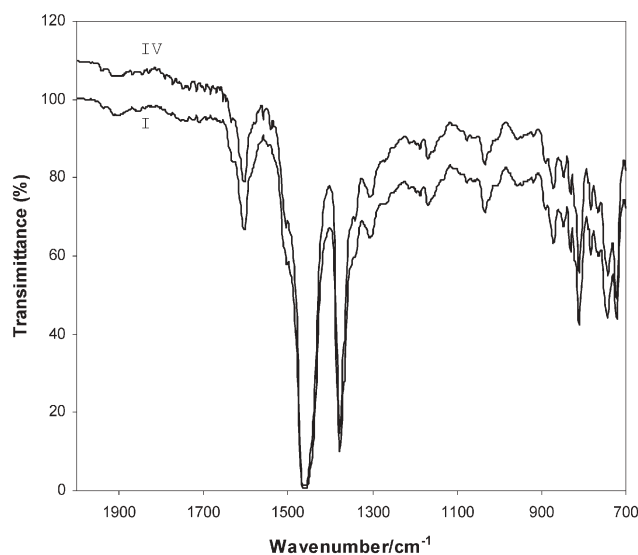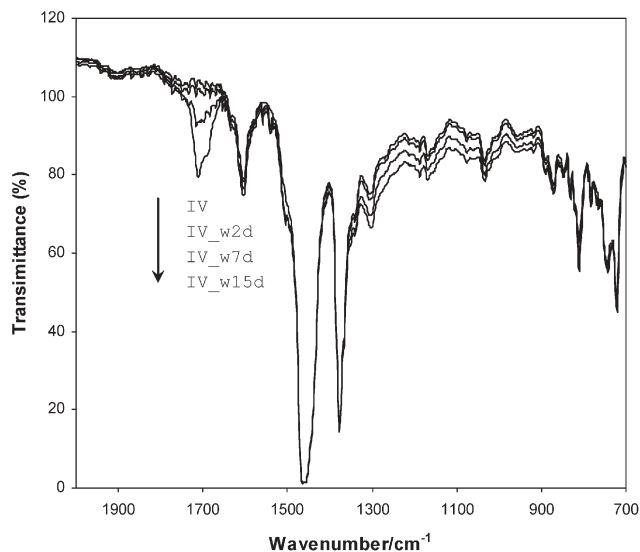
Fig. 4 Average FTIR spectra of at least ten replicates of fresh and weathered samples of diesel IV. IV: fresh, IV_w2d: weathered 2 days, IV_w7d: weathered 7 days, IV_w15d: weathered 15 days.

In order to investigate the performance of each numerical method, three series of comparisons for each oil sample were carried out. There were two series between a fresh sample and (a) its own weathered derivatives, and (b) weathered derivatives of the other sample. The third series consisted of comparisons between all weathered samples of the same fresh sample. The comparison of two fresh samples was carried out as a reference. The comparisons in the third series provided different weathering times from two to 15 days, which started from different stages—not all from a fresh sample.

The results, (Table 2) show the general changes that accompany the development of weathering. Due to the similarity of the samples which were all diesels and had been mildly "weathered" for no more than 15 days, quite a few indices are high and distribute in a narrow range. This is more obvious in the generic range in which very high similarities can be found because the peaks in this region are wide, modest and not specific for a particular diesel. From the results, it is obvious that sample IV changes more than sample I during weathering. Generally speaking that all methods work well in fingerprint and generic regions, the greater the weathering time difference between samples, the smaller the similarity indices. It is noted, however, *e.g.* that the scores of sample IV in the fingerprint region have a turn-over point where the 7-day weathered samples (weathered 7 days from fresh sample) show a greater change than the 8-day (weathered 8 days from a sample already weathered 7 days) weathered samples, which suggests a slow-down of weathering in later development. This demonstrated that the effects of weathering are less at longer times.

The next investigation was carried out by using the median similarity between fresh samples of I and of IV as the threshold of match regardless of weathering degree—any similarity that has a value smaller than the threshold in the fingerprint and generic regions indicates that the two samples being compared do not match. (Note that in the supplementary region, because

Fig. 3 Average FTIR spectra of samples of diesel oils I and IV from measurements on ten independent samples.

Table 2 Median scores of the comparisons of the FTIR spectra of weathered samples of I and IV[a]

	Fingerprint region				Generic region			
	1stCor	Cor	1stCos	Cos	1stCor	Cor	1stCos	Cos
I-I_w1d	0.992	0.996	0.996	1.0000	0.997	0.999	0.998	1.0000
I_w2d-I_w7d	0.994	0.996	0.997	1.0000	0.995	0.998	0.998	1.0000
I-I_w7d	0.976	0.986	0.988	0.9999	0.989	0.997	0.995	1.0000
I_w7d-I_w15d	0.973	0.984	0.986	0.9998	0.953	0.970	0.977	0.9999
I_w2d-I_w15d	0.944	0.967	0.972	0.9997	0.931	0.960	0.966	0.9999
I-I_w15d	0.911	0.948	0.954	0.9995	0.922	0.955	0.961	0.9998
IV-IV_w1d	0.980	0.993	0.990	0.9999	0.993	0.999	0.997	1.0000
IV_w2d-IV_w7d	0.955	0.977	0.977	0.9998	0.973	0.986	0.987	0.9999
IV-IV_w7d	0.901	0.952	0.949	0.9995	0.956	0.980	0.978	0.9999
IV_w7d-IV_w15d	0.949	0.968	0.974	0.9997	0.949	0.970	0.975	0.9999
IV_w2d-IV_w15d	0.830	0.898	0.911	0.9991	0.884	0.925	0.941	0.9997
IV-IV_w15d	0.761	0.862	0.872	0.9987	0.865	0.913	0.931	0.9997
I-IV	0.962	0.914	0.981	0.9992	0.945	0.987	0.973	1.0000
I-IV_w2d	0.959	0.913	0.979	0.9992	0.942	0.986	0.971	1.0000
I-IV_w7d	0.878	0.841	0.937	0.9986	0.899	0.961	0.950	0.9999
I-IV_w15d	0.718	0.706	0.847	0.9973	0.828	0.893	0.912	0.9996
IV-I_w2d	0.946	0.907	0.972	0.9991	0.935	0.986	0.968	1.0000
IV-I_w7d	0.926	0.904	0.962	0.9991	0.934	0.988	0.967	1.0000
IV-I_w15d	0.871	0.888	0.933	0.9988	0.886	0.956	0.943	0.9999

[a] Note: A different presentation accuracy of the results of Euclidean cosine squared is used for revealing the differences between them.

of the nature of the similarity index, a value *smaller* than the threshold means there is a match.) The similarity between samples I and IV were used as a threshold of match/no match. If a pair of spectra derived from the same fresh sample is mis-classified as no-match then a type I error (false negative) has been made, and any error occurring in the comparison of a fresh sample with weathered samples of the other fresh sample is type II error (false positive). The results of comparing these spectra are shown in Table 3.

In the fingerprint region, correlation coefficient squared and Euclidean cosine squared can correctly discriminate 10 out of 12 relationships (two type I errors) but the latter has two type II errors. It can be seen that Euclidean cosine squared has a very narrow distribution of similarity indices which could limit its use. The results of the first difference methods are very poor, with less than half the spectra being correctly classified. Using the information from the generic region the two first difference-based methods gave better results then other methods. This result is not surprising because in the fingerprint region there are more and sharper peaks then in generic region. First difference methods are more sensitive to any changes because of differentials of successive intensities. In the supplementary region, using the absolute difference between averages, the number of type I errors is high: 10 out of 12 are incorrect. Because of its low correct discrimination rate, this region can only be used to test the occurrence of weathering

and provides limited supplementary information about the relationship between two spectra.

The error rate of the comparisons related to sample I is smaller than that of the comparisons related to sample IV. The differences between the members of sample IV with different degrees of weathering are greater, which indicates weathering has had a greater effect on sample IV than on sample I. In addition the distributions of the comparison results related to sample IV are wider than their counterparts of sample I. In most cases, the number of type I errors (false negative) is greater than the number of type II errors (false positive) because longer weathering changes spectra significantly compared to the references. That is to say, although weathering makes it harder to conclude that a sample is from a particular (not only fresh) source, weathered samples from different sources are still mostly correctly classified. Samples of the same oil weathered to some extent can be matched, therefore we propose that artificial weathering of fresh oils could help the analyst to match an environmental sample to its source from very similar suspects.

The ROC curves for the comparisons of weathered samples are shown in Fig. 5 and Fig. 6. There are fewer steps in the curves than many other ROC curves found in the literature. When the thresholds are changed, the true positive rates and the false positive rates do not change simultaneously. It is apparent that when weathering has happened the true positive

Table 3 Discrimination errors of methods over 12 comparisons between oil samples I and IV with different extents of weathering. Type I errors are the mis-classification of the same oils (albeit weathered) as 'not-matched'. Type II errors are the mis-classification of different oils as 'matched'. The figures in parentheses break the errors down for sample I, sample IV

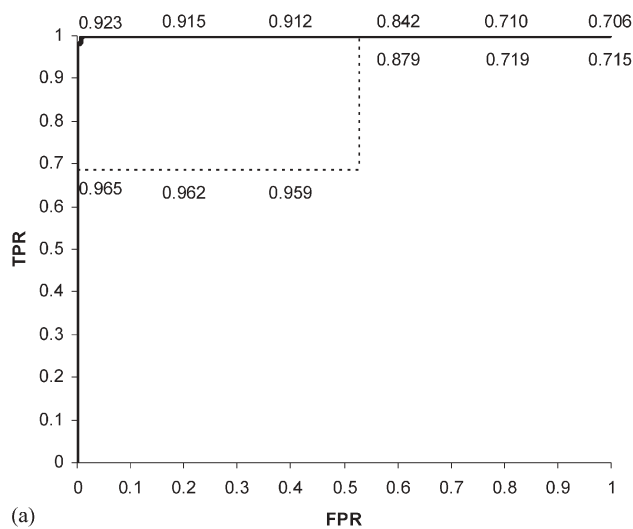| Method	Fingerprint Region		Generic Region		Supplementary Region	
	Type I	Type II	Type I	Type II	Type I	Type II
1stCor	7 (2, 5)	0	2 (0, 2)	2 (0, 2)	10 (5, 5)	2 (0, 2)
Cor	2 (0, 2)	0	8 (3, 5)	1 (0, 1)		
1stCos	7 (2, 5)	0	2 (0, 2)	2 (0, 2)		
Cos	2 (1, 1)	2 (0, 2)	3 (1, 2)	4 (1, 3)		
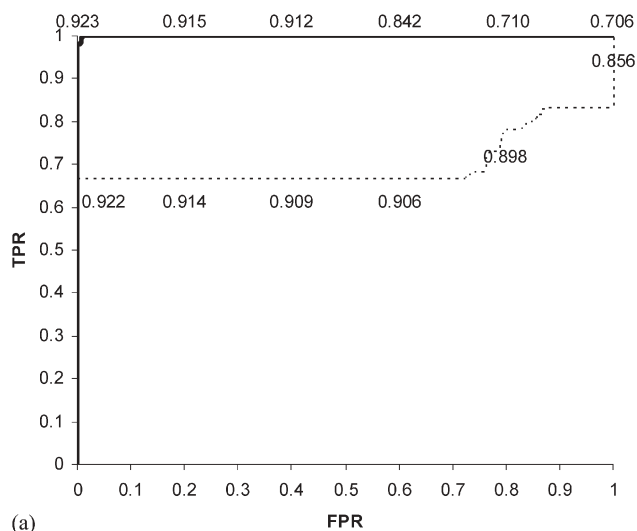
(a)
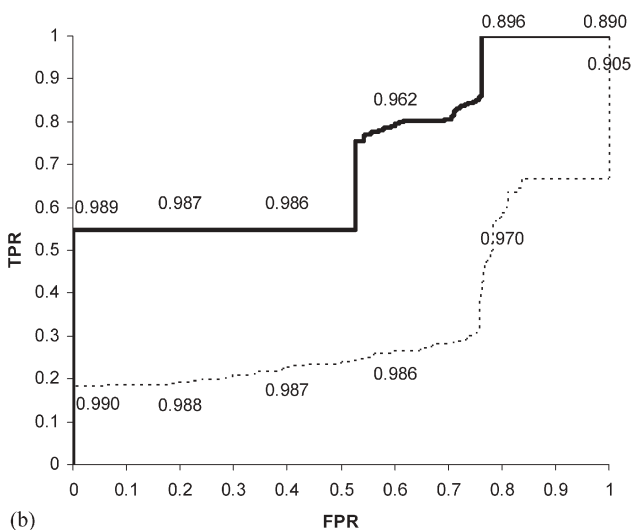
(b)

Fig. 5 Receiver operating characteristic (ROC) curves of decisions on matching using Cor between a fresh sample and its weathered derivatives ("matched" set) against another fresh oil and its weathered derivatives ("non-matched") in (a) fingerprint and (b) generic regions. The numbers shown along the lines are decision indices for two samples respectively. Solid line: sample I as the reference of "matched" set; dotted line: sample IV as the reference of "matched" set.

rates are less than 1, and they marginally increase when the lower matching thresholds are applied. Compared to the weathering effects on the samples, sampling and measurement uncertainties of samples with the same weathering degree are much smaller. So most of the changes are abrupt rather than smooth.

In agreement with the inference that can be drawn from Table 3, the two ROC curves of correlation coefficient squared (Fig. 5), which were created using samples I and samples IV, respectively, as the references of the "matched" set, reveal that samples I and IV respond to weathering differently. It can be seen that, *e.g.*, in the fingerprint region, if 0.922 is chosen as a cut-off point all weathered samples can be categorised to their real sources without any mis-classification. But sample I has a TPR close to 1 and sample IV has a TPR of only about 0.67.

(a)
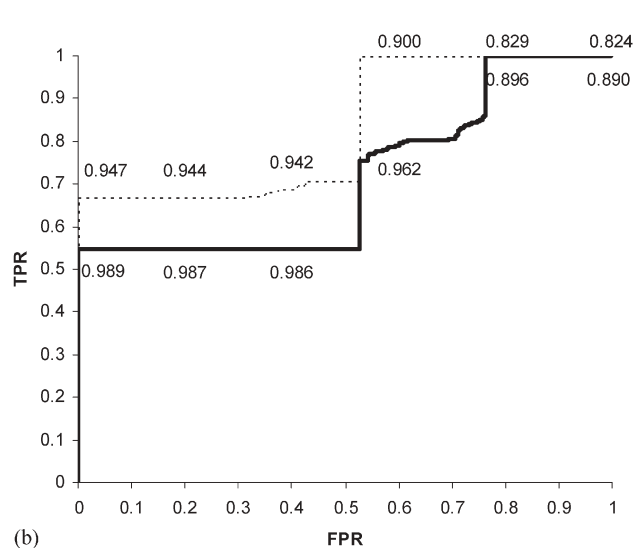
(b)

Fig. 6 ROC curves of Cor and 1stCor of comparisons using sample I as the reference of "matched" set in (a) fingerprint and (b) generic regions. The numbers shown along the lines are decision indices for two samples respectively. Solid line: Cor; dotted line: 1stCor.

That is to say that if the similarity index is not less than 0.922 all weathered samples of sample I can be classified correctly without introducing any false positive classification but sample IV has 33% weathered samples that cannot be linked back to their source if no false positive classification is allowed. When lower scores are used, the correct classification rate does not increase much but more false positive classification occurs. In the generic region, TPRs are lower and FPRs are higher, where most of the ROC curve of sample IV is below the diagonal, which means the situation is worse than in the fingerprint region, but the weathered samples of sample I can be classified more easily than those of sample IV. Based on the similarity scores, TPR and FPR, it can be seen that using correlation coefficient squared it is possible to match samples that have been weathered for 7 days to fresh samples based on the information from the fingerprint region.

The performances of correlation coefficient squared and first difference correlation coefficient squared were compared

using ROC curves, with the spectra of sample I as the reference of the "matched" set, and the result is shown in Fig. 6. We can see that the different matching methods work differently in different spectral regions. In the fingerprint region, the correlation coefficient square can achieve 0.98 TPR with zero FPR, while first difference correlation coefficient square has to accept 0.5 FPR to get an equivalently high TPR. But in the generic region, even though neither can give very certain classifications (close to 1 TPR with zero FPR), the first difference method does achieve a higher TPR when it is compared to a straight correlation coefficient squared at the same FPR.

Discussion

An oil is a mixture with a complex composition which leads to an IR spectrum with many convoluted peaks that cannot be resolved. In this sense, although FTIR is a sensitive technique it lacks the clear resolution of NMR or GC coupled with mass spectrometry. A non-zero baseline is also common in oil infrared spectra which may be likened to the so-called UCM (unresolved complex mixture)[20] in a gas chromatogram. Some environmental protection agencies use GC-MS routinely, but the cost and time involved do recommend FTIR (and fluorescence) as methods for at least preliminary screening. A numerical procedure will add value to the use of these methods.

A spilled oil sample coming from the real world comes with other material and a sampling technique for withdrawing the oil without water and debris is essential. Weathering introduces other effects on the spilled oil which cannot be easily predicted. Suspect sources are always very similar and might even come from the same batch of crude. Normally only parts of the spectrum of a spilled oil can clearly match the suspect source. The use of ASTM D3414 depends on the experience of the analyst and different analysts may have different judgements when they are looking at the same spectra. Under these circumstances, numerical comparison methods are quicker, easier and more objective than the traditional overlay procedure. This is not to say that numerical methods should always replace the role of an analyst.

Spectra do not change linearly, completely and synchronously to the change of chemical composition. Different numerical spectral comparison methods have their characteristics. Correlation coefficient squared and Euclidean cosine squared check linear relation between two series of data. The difference derivatives are based on the difference of adjacent intensities and are sensitive to smaller differences between spectra. In general situations, if samples are not too similar and not weathered, in which case spectral differences between samples exist in all interesting spectral regions, first difference-based similarity statistics work better than the parent measures. For weathered oils, there are always some changes in the fingerprint region even though they are smaller than the changes in the generic region. When weathering happens and two samples are quite similar, it is difficult to match a severely weathered sample with its fresh form without any false positive classification. In order to establish the link between weathered oils and their fresh source, correlation coefficient squared

should be used in the fingerprint region and first difference methods should be used to improve checking the degree of weathering. We conclude that it is not possible to reliably match a sample weathered for more than 7 days to its source. However, for a sample weathered for a short time it is possible to match it to its real source even among quite similar potential suspects. Different similarity indices should be used in different regions of the spectrum.

Even if it has negative impacts on matching, weathering is a potential source of information in making a choice between two very similar suspects. As we found, the effects of the simple weathering we used here depend on the properties of the oil and the weathering patterns are different for different oils. The difference between a weathered sample and its source is smaller than the difference between this weathered sample and any unrelated sample, and this difference will increase with the degree of weathering. If the spill is weathered and suspect sources are quite similar, artificial weathering is a good, and maybe the only, way to indicate the real source. When artificial weathering is used in matching analysis, it is not necessary to reproduce the exact weathering processes which occurred with the spilled oil. The results shown here do not cover all situations involved in weathering because weathering is a very complex process, and here only mainly photooxidation and evaporation happened. We have not tried to set up a library to cover all weathering processes but tried to provide a generic simple method for testing matches between a spill and possible sources.

As shown in this paper, a receiver operating characteristic curve is a useful decision making tool which can not only give more flexible choices depending on different situations when it is compared to the single point decision, but also it gives the possibility of a decision based on the available knowledge. In a complex situation, like matching weathered oils with their fresh source in which errors always exist, a ROC curve will help finding a balanced decision point where true suspects can be caught with the maximal acceptable errors.

Acknowledgements

References

1 J. W. Short and A. Heintz, *Environ. Sci. Technol.*, 1997, **31**, 2375–2384.
2 U. S. EPA, U.S. EPA Oil Program, http://www.epa.gov/oilspill/index.htm, accessed: October 2004.
3 J. S. Mattson, Workshop on Pattern Recognition – Applications to Oil Identification, NOAA, Washington, 1977, p. 113.
4 J. Li, D. B. Hibbert, S. Fuller, J. Cattle and C. Pang Way, *Anal. Chem.*, 2005, **77**, 639–644.
5 N. Dimov and A. Pavlova, *J. Environ. Monit.*, 2000, **2**, 266–270.
6 D. Staniloae, B. Petrescu and C. Patroescu, *Environ. Forensics*, 2001, **2**, 363–366.
7 J. Li, S. Fuller, J. Cattle, C. Pang Way and D. B. Hibbert, *Anal. Chim. Acta*, 2004, **514**, 51–56.
8 J. H. Christensen, A. B. Hansen, G. Tomasi, J. Mortensen and O. Andersen, *Environ. Sci. Technol.*, 2004, **38**, 2912–2918.
9 M. J. Wade, *Preprints of Extended Abstracts presented at the ACS National Meeting*, American Chemical Society, Division of Environmental Chemistry, 1998, vol. 38, pp. 236–238.

10 ASTM, *D 3414-98 Standard test method for comparison of waterborne petroleum oils by infrared spectroscopy*, American Society for Testing and Materials, Philadelphia, 1998.

11 M. Reed, O. Johansen, P. J. Brandvik, P. Daling, A. Lewis, R. Fiocco, D. Mackay and R. Prentki, *Spill Sci. Technol. Bull.*, 1999, **5**, 3–16.

12 W. Lehr, R. Jones, M. Evans, D. Simecek–Beatty and R. Overstreet, *Environ. Modell. Software*, 2002, **17**, 191–199.

13 M. F. Fingas, *J. Hazard Mater.*, 2004, **107**, 27–36.

14 K. A. Burns, S. Codi, C. Pratt and N. C. Duke, *Org. Geochem.*, 1999, **30**, 1273–1286.

15 D. E. Nicodem, C. L. B. Guedes, M. C. Z. Fernandes, D. Severino, R. J. Correa, M. C. Coutinho and J. Silva, *Prog. React. Kinet. Mech.*, 2001, **26**, 219–238.

16 P. C. Bennett, F. K. Hiebert and J. R. Rogers, *Hydrogeol. J.*, 2000, **8**, 47–62.

17 C. D. Brown and H. T. Davis, *Chemom. Intell. Lab. Syst.*, 2006, **80**, 24–38.

18 M. S. Pepe, *Biometrics*, 2000, **56**, 352–359.

19 X.-H. Zhou, D. K. McClish and N. A. Obuchowski, *Statistical methods in diagnostic medicine*, Wiley, New York, 2002.

20 M. A. Gough and S. J. Rowland, *Nature*, 1990, **344**(6267), 648–650.