Multivariate statistical approach to the fingerprinting of oils by infrared spectrometry

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 1977

Impact Factor: 5.64 · DOI: 10.1021/ac50010a029

Reprinted from **ANALYTICAL CHEMISTRY**, Vol. 49, Page 297, January 1977

Multivariate Statistical Approach to the Fingerprinting of Oils by Infrared Spectrometry

James S. Mattson,*[1] Carol S. Mattson, Mary Jo Spencer, and Scott A. Starks[2]

Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, Fla. 33149

The need to attach a statistically significant number, describing the extent of match or mismatch of infrared spectra of oil samples, is satisfied using a multivariate normal probability density function. The requirements of normality and independence are examined, and two methods of baseline placement are considered. Example results are presented for a test population of 10 lubricating oils, 22 waste lubes, 30 No. 2, 30 diesel, 12 No. 4, 10 No. 5, 28 No. 6 fuels, and 62 crude oils.

The problem of identifying the source of a clandestine oil spill has been approached in a myriad of ways. At one time, it was seriously suggested that the concept known as "active tagging" be pursued. Active tagging involves the addition of some uniquely coded foreign material to every water-borne cargo of crude oil and refined fuel. The technological problems of such proposals are sufficiently difficult to warrant skepticism, but the costs of implementing an active tagging scheme are merely astronomical.

The alternative to active tagging has been referred to as "passive tagging", involving the measurement of a sufficient number of properties of an oil to uniquely distinguish it from any other oil. Several analytical techniques have been proposed for the identification of oils. Besides infrared spectrometry (*1–3*), these include gas chromatography (*4, 5*), neutron activation analysis (*6, 7*), mass spectrometry (*8*), fluorescence spectrometry (*9*), and thin-layer chromatography (*10*). All of the methods proposed suffer from one major defect, the lack of a clearly defined metric which will accurately predict the probability that two sets of measurements ("patterns") represent the same oil. A library of such patterns, whether it consists of trace element concentration patterns obtained from neutron activation analysis or infrared absorption intensities, is amenable to statistical analysis using classical multivariate statistics or to more sophisticated, nonparametric pattern recognition techniques. This paper is addressed to the problem of calculating as accurately as possible the probability that two infrared spectral patterns which exhibit modest differences do not do so as a result of analytical error alone.

The envisioned application of the metric arrived at in the fashion described in this paper is described as follows. Suppose there are ten possible suspects, any one of which may have been responsible for a clandestine oil spill, and that a sample is available from each of the suspects as well as the spill. If there was no doubt that one of the ten was the guilty party, then it would be trivial to determine which one, provided that all ten suspect patterns are unique. For this trivial case, one could compute the Euclidean distance between the spill pattern and each suspect pattern in n-space (a pattern is just the n-tuple $(x_1, x_2, \ldots, x_n)$), and then associate guilt with the suspect pattern that yields the smallest such distance (i.e., the "closest" match). This can sometimes be done using nature's own pattern recognition device; i.e., the "eyeball" approach. It is not reasonable, however, to assume a priori that the guilty party has been sampled. If there is a chance that the guilty party may not have been sampled, the above approach is invalid and it is necessary to assign a reasonable probability to each spill-suspect match based upon our knowledge of 1) the precision of the analytical method, and 2) the distribution function for the patterns over the population of all oils, not just over those few samples which make up the suspect pool. This paper is an effort to define a probability algorithm which could be modified to apply to any analytical procedure proposed for use in oil identification.

The primary considerations in this effort should be applicable to other analytical methods, and are summarized as follows:

1) Determine the probability density function for each variable, over a sufficiently large and random population of oils. If the variables do not obey a normal (Gaussian) distribution, normal multivariate statistical techniques are invalid.

2) Determine the analytical variance for each variable. Compare it to the population variance. If the two are of the same order of magnitude, the usefulness of the variable is limited.

3) Examine the independence of each variable from the others. If the variables are interdependent to a high degree, the calculation of a probability can be difficult.

EXPERIMENTAL

Transmission infrared spectra of 204 samples of oil were obtained using both precision sealed and demountable KBr cells (Wilks Scientific Corp., South Norwalk, Conn.), with cell pathlengths from 0.09 to 0.15 mm. The spectra were obtained with a Data General NOVA computer–Perkin Elmer 180 spectrometer system which has been described elsewhere (*11*), recording spectra from 2000 to 650 cm^{-1} in 1-cm^{-1} intervals, with a spectral slit width of 1.0 ± 0.2 cm^{-1} and an ordinate precision of ca. 1.2 × 10^{-3} absorbance unit at 1800 cm^{-1}. The spectra were smoothed with a 21-point quartic smooth (*12*),

[1] Present address, Center for Experiment Design and Data Analysis, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, 3300 Whitehaven St., N.W., Washington, D.C. 20235.

[2] Present address, Department of Electrical Engineering, Rice University, Houston, Texas.

Table I. Infrared Absorption Bands Used for Pattern Recognition Studies of Petroleum Oils

	Kawahara (1), cm^{-1}	Mattson (2),[b] $cm^{-1} \pm 1s_p$	Lynch and Brown (3), $cm^{-1} \pm 2.5\ cm^{-1}$	This study,[b] $cm^{-1} \pm 1s_p$
1	...	1694 ± 7.5	...	...
2	...	...	...	1629 ± 1.9
3	1600	1600 ± 3.3	...	1603 ± 5.6
4	...	...	...	1518 ± 2.0
5	1460	1456 ± 3.2	...	1456 ± 5.4
6	1375	(1375)[a]	...	1376 ± 1.4
7	...	(1309)[a]	...	1304 ± 1.2
8	...	1168 ± 6.8	1160	1166 ± 2.0
9	...	...	1145	1154 ± 2.0
10	...	...	1070	...
11	1027	1034 ± 2.8	1020	1032 ± 1.0
12	...	...	955	963 ± 2.9
13	...	...	915	918 ± 0.7
14	...	...	890	888 ± 1.1
15	870	874 ± 3.4	870	870 ± 1.4
16	...	...	845	846 ± 1.4
17	...	...	835	832 ± 0.6
18	...	...	820	...
19	810	814 ± 2.6	805, 810	809 ± 1.6
20	...	...	790	793 ± 1.6
21	...	...	780	781 ± 2.0
22	...	...	765, 770	765 ± 1.8
23	...	747 ± 2.8	740	741 ± 2.4
24	720	725 ± 2.6	720, 725	722 ± 1.5
25	...	...	695	697 ± 1.7
26	...	...	...	673 ± 1.1
Total	7	11	18 (21)	23

[a] Peaks not used in fingerprints. [b] s_p is defined as the "estimated standard deviation" in the position of the peak maximum, measured from all spectra in which the peak exhibited a distinct maximum.

normalized to a 0.10-mm pathlength, and stored on cassette tapes.

Of the 204 samples, 182 were (fresh and unweathered) lubricating oils (10), No. 2 fuels (30), diesel fuels (30), No. 4 fuels (12), No. 5 fuels (10), No. 6 fuels (28) and crude oils (62). The other 22 samples were waste automotive crankcase lubricants, obtained from service stations in the Miami, Fla., area. All samples were stored at 5 °C from the time of receipt to the time of analysis. The 182 fresh samples were analyzed neat; the 22 waste lubes were centrifuged for 1 h at about 35 °C to remove water and particulate matter.

Replicate analyses, carried out to determine the analytical variances for each variable, consisted of six analyses each of a No. 6 fuel, a No. 2 fuel, and a Kuwait crude. These three samples were American Petroleum Institute "pool" samples, obtained from the Department of Biology, Texas A&M University.

RESULTS AND DISCUSSION

Variable and Baseline Selection. Table I lists frequencies of infrared absorption bands used in previously reported oil identification studies. Kawahara (1) employed the seven bands listed in Table I in various nonlinear combinations, conveniently eliminating differences due to variations in sample thickness. The use of variable ratios reduces the overall number of variables by one, reducing the dimensionality used by Kawahara (1). Mattson (2) measured absorption band *areas* using a computer–spectrometer system (13), normalizing each to the area of the 1456 cm^{-1} peak, also to eliminate pathlength as an uncontrolled variable. Of eleven peaks encoded by Mattson (2), only eight were employed in the final patterns. One was lost in the normalization process, one deleted due to high analytical error (1309 cm^{-1}), and the 1375 cm^{-1} peak was dropped due to its high correlation with the 1456 cm^{-1} peak (rank correlation coefficient, r_s (1456, 1375) = 0.97).

In their more recent study, Lynch and Brown (3) manually encoded the absorptivities for the 21 bands listed in Table I and stored them in a computer file. In order to eliminate sample pathlength as a variable, they (3) compute the average ratio (more recently, log ratio (14)) for the corresponding peaks in a pair of spectra, and then use the differences from the mean ratio as a measure of similarity.

In this study, 23 bands were initially chosen for automated encoding. These are listed in the last column of Table I. A search is conducted for each peak within a 9-cm^{-1} region defined as the nominal peak position ±4 cm^{-1}. If no local maximum is encountered within that 9-cm^{-1} range (i.e., greater than either of the points just adjacent to the 9 cm^{-1} being searched), the absorbance at the nominal frequency is used as the value for that variable. A careful study was made of two baseline techniques, the tangent-line technique used by Kawahara (1), Mattson (2), and Spencer (15) and a flat baseline method similar to that used by Lynch and Brown (3). Figure 1 illustrates the two baseline approaches for two peaks in a No. 2 fuel oil. The difference between the two measurements is indicated by A_u, the "unresolved", or background, absorption. A_u is included in the flat baseline measurement but excluded when a tangent line is used. The use of a tangent-line technique would cause a zero to be recorded at each of those points where no maximum was encountered within ±4 cm^{-1} of the nominal position. This would result in a plethora of pseudo-identical values in many of the patterns. By including the unresolved "background" absorption in the measurement of each variable, this problem is eliminated.

In order to use the flat baseline approach, a single point was chosen as the reference zero for the entire spectrum. Because of the lack of water vapor absorption in the area, and the fact that no oil exhibited a band nearby, 1990 cm^{-1} was chosen for the baseline value. The effect of using a baseline zero at other than 0 absorbance is to correct for reflectance losses caused by the cell windows and changes in the 100%T adjustment of the spectrophotometer.

In a study of a 72-oil subset of the 204-oil library reported in this paper, Spencer (15) employed the tangent-line baseline

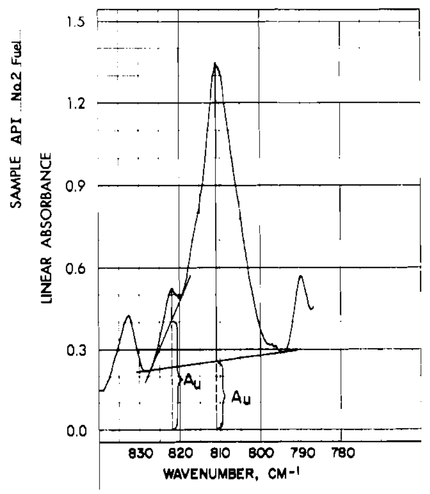

Figure 1. An illustration of the tangent-line and flat baseline approaches to measuring the height of absorption bands

Table II. Comparison of Baseline Methods for Oil Spectra

	Estimated standard deviation, s, in absorbance units × 10^3			
	Analytical[a]		Population[b]	
Peak cm^{-1}	Tangent	Flat	Tangent	Flat
1603	8.^{2}	6.^{0}	238.^{3}	303.^{9}
1304	12.^{3}	10.^{1}	43.^{7}	191.^{1}
1166	12.^{8}	8.^{7}	35.^{0}	121.^{2}
1032	10.^{5}	10.^{0}	25.^{7}	136.^{9}
963	5.^{1}	4.^{0}	63.^{9}	72.^{4}
846	4.^{9}	7.^{4}	38.^{7}	223.^{4}
809	14.^{3}	9.^{7}	211.^{9}	435.^{4}
765	11.^{9}	9.^{6}	76.^{9}	240.^{6}
741	40.^{3}	28.^{4}	228.^{6}	357.^{4}
722	18.^{7}	8.^{7}	205.^{6}	148.^{6}
697	8.^{3}	3.^{2}	101.^{1}	129.^{1}
[c] Total s	54.^{1}	39.^{9}	470.^{8}	794.^{6}

[a] Based on 6 replicates each of a No. 2, a No. 6, and a Kuwait crude. [b] Based upon spectra of 26 crude oils, 10 lubes, 10 No. 2, 8 No. 4, 8 No. 5, and 10 No. 6 fuel oils. [c] Total $s \equiv (\Sigma s^2)^{1/2}$

Table III. Frequency of Occurrence and Analysis of Variance for 23 Infrared Absorption Bands

		Relative s, %		
Peak, cm^{-1}	Occurrence, %	Analytical	Population	Ratio s_p^2/s_a^2
1629	22.1	1.04	75.62	7,240
1603	98.0	0.94	50.49	2,885
1518	8.8	3.38	69.58	424
1456	100.0	13.39	15.25	1.3
1376	100.0	10.86	10.04	0.9
1304	86.3	1.43	29.11	414
1166	73.5	1.88	41.28	482
1154	53.4	3.04	43.94	209
1032	96.1	2.17	34.16	248
963	65.2	1.10	22.46	417
918	74.5	1.89	19.96	112
888	38.2	1.05	43.35	1,705
870	86.3	1.10	53.53	2,368
846	81.9	1.42	46.18	1,058
832	41.7	0.58	56.78	9,584
809	94.1	1.59	53.66	1,139
793	5.4	0.98	47.44	2,343
781	64.2	1.54	42.21	751
765	72.5	1.28	40.91	1,021
741	83.3	2.92	39.56	184
722	95.6	1.07	17.41	265
697	70.6	0.85	36.41	1,886
673	71.6	1.60	106.00	4,389

method to classify and identify oils. After her study was completed, the flat baseline approach was applied to the 72-oil subset to determine which method should be applied to oil identification. The first conclusion was an obvious one. The flat baseline technique should yield better analytical precision than the tangent-line approach, as only two measurements are made that are subject to random error; the absorbance at the peak maximum and the absorbance at the frequency used for establishing the baseline. The tangent-line approach has three such sources of random error. The predicted ratio of the analytical variance of the tangent-line technique to that of the flat baseline is 1.50. This is close enough to the measured value of 1.84 to warrant consideration of the flat baseline technique on analytical precision grounds alone. The reason that the variance ratio exceeds 1.50 by 23% probably lies in the fact that one can choose the baseline reference point in a noise-free region of the spectrum, where energy throughput is high and atmospheric water vapor absorption is low. In this paper, the reference point was 1990 cm^{-1}, and the analytical standard deviation measured at 1990 cm^{-1} was only 1.2 × 10^{-3} absorbance unit.

The preceding point requires consideration when considering any analytical technique, but another one, "information content", is unique to the type of analytical problem being considered here. Since differences between infrared spectra of some oils are so subtle as to be discernible to only the most experienced spectroscopists, sophisticated mathematical techniques which fall under the general heading of "pattern recognition" often must be employed to make the final analysis. These mathematical methods, which include the classical one of linear discriminant function analysis, may make simple decisions with only minimal information, but when the decisions begin to tax nature's own pattern recognition system, the human sensory system, it is essential that the analytical data input to the decision-making process be as complete as possible.

Two important questions are: i) is there additional information in A_u, shown in Figure 1, and ii) if so, can it improve the results in identification? In Table II, the "spread" of peak absorbance values for the 72-oil subset, s_p, is represented by the square root of the estimated population variance, peak-by-peak, for eleven major oil identification peaks. In every case, the greater s_p for the flat baseline method means that there exists a greater diversity of peak heights when A_u is included. When one examines the ratio of the estimated population variance to the estimated analytical variance, (s_p^2/s_a^2), the ratio is about 400:1 for the flat baseline approach and is only 75:1 for the tangent-line scheme. Thus the combination of increased population variance with decreased analytical error gives the flat baseline approach *five times the information content* of the tangent-line method. This point is directly applicable to fingerprinting, where the most important aspect of the analytical procedure is to emphasize differences between oils while minimizing the chance of an accidental match due to random error.

Table III is intended to complete the "information content" discussion, and is confined to the flat baseline results, using all 204 oils, and the 23 peaks listed in Table I. The column labeled "occurrence" refers to the percentage of the 204 patterns that exhibited distinct maxima in the 9-cm^{-1} range about each nominal peak position. The estimated analytical standard deviations, s_a, were computed using the eighteen replicate analyses described in the experimental section, while the population values originate from the 204-oil library (with the exception of the 1456 and 1376 cm^{-1} s_p values, which came *from the 72-oil subset).* The large analytical errors and low population ranges for the 1456 and 1376 cm^{-1} peaks justified the deletion of these two peaks from any further consideration in fingerprinting, since any difference in absorbance values in these two variables can be accounted for between two oils by analytical error alone.

Testing Normality. The prerequisite for employing multivariate normal statistics is that each variable exhibit a normal (Gaussian) distribution. Since the authors intend to use these data for oil classification as well as identification, it is important that the population density function be examined for any deviations from normality. The approach taken to evaluate the population was as follows. The original 72-oil subset of the population was assumed to be randomly chosen, which it was within the confines of a selection procedure that attempted to collect the widest possible variety of crude oils and refined products. The estimated means, m_i, and standard deviations, s_i, for the 72-oil population were computed, as were the statistical parameters of skewness and Kurtosis. For 72 patterns, the variables exhibited near-normal parameters. On scaling the populations up to 182 oils (prior to the addition of the 22 waste oils to the library), the values for m_i and s_i ($i = 1, 2, \ldots, 21$) were recomputed (the 1376 and 1456 peaks had been deleted). This was done with the assumption that if the m_i and s_i values remained unchanged on scaling up to 182 patterns, then the original 72 oils must have adequately represented the population of all oils and it would be unnecessary to carry the "normalcy" test beyond 182 patterns. In fact, the rms (root mean square) change in the m_i values was only 4.4% and the rms change in the s_i values was 9.1%. Since the estimated means and standard deviations for all 21 variables exhibited such modest changes on increasing the number of randomly selected oils from 72 to 182, and the Gaussian character of the 182-pattern library was better than that of the 72-pattern library, the authors believe that it is reasonable to postulate that the infrared spectra of oils exhibit normal statistical behavior, and also to assume that the present 204-pattern library adequately represents the entire population of crude oils and refine products.

The Multivariate Normal Algorithm. For n dimensions, the multivariate probability distributions define hyperellipsoids (Equation 1) about the point X_0 of constant probability

$$\chi^2 = (X_0 - X_1)'M^{-1}(X_0 - X_1) \tag{1}$$

that are related to the density function ϕ (χ^2) (Equation 2) (*16*),

$$\phi(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} e^{-\chi^2/2}(\chi^2)^{\nu/2-1} \tag{2}$$

where ν equals degrees of freedom, M^{-1} is the inverse of the variance-covariance matrix with elements $m_{ij} = \rho_{ij}\sigma_i\sigma_j$, where ρ_{ij} is the correlation coefficient between variables i and j and σ_i and σ_j are the standard deviations of the ith and jth variables. Thus the diagonal elements of M are just the variances of the n variables, σ_i^2. If two patterns in n-space are to be tested for possible membership in the same class (i.e., originating from the same oil), and the only source of difference between the patterns is random analytical error in the measurement process, then several simplifications result. Since the differences in each variable $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$ are assumed to be due to random measurement errors, the *errors* can be assumed to be independent, and the off-diagonal elements of M set equal to zero. In addition, the diagonal elements σ_i^2 can be approximated by the measured values s_i^2 given in Table III. A further assumption in such a case is to consider that the error vector $\Delta X \equiv (\Delta x_1, \Delta x_2, \ldots, \Delta x_n)$ is multivariate normal. In that case, the metric computed with Equation 1 is distributed as χ^2 with n degrees of freedom. This allows the calculation of the probability that two patterns are a "match" using ordinary univariate tables of χ^2 (*16*).

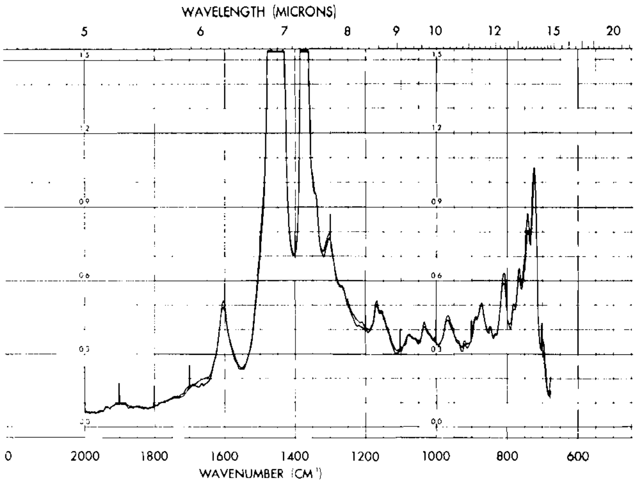

Figure 2. Infrared spectra of Libyan and Julesburg, Colo., crude oils, normalized to 0.1-mm pathlength, set equal at 1990 cm^{-1}, and 2X ordinate expansion. 99.9% of all possible pairs in the 204-oil library exhibit less similarity than this pair

For the simplest case, where the errors associated with each variable are assumed to be independent, Equation 1 reduces to the simple expression shown in Equation 3 (*17*).

$$\chi^2 = \sum_{i=1}^{n} (\Delta x_i/s_i)^2 \tag{3}$$

Equation 3 is rigorously valid only when the absolute values for the variances are known.

In the comparison of absorption spectra of two samples of oil using 21 variables (peak absorption magnitudes at 21 different frequencies), a Chi-square of 40 indicates a probability of only 0.01 that a difference this great or greater could be expected from random analytical error alone, *assuming that the two samples are in fact identical.* Because of the complexities of infrared spectra of oils, and the realization that sampling and instrumental errors are not insignificant, it is desirable to assess the variations which can be expected to occur as a matter of course when pairs of oils are *known to be different.*

Making the assumption that two oils are identical (though known to be different), and examining the 20,706 possible pairs of patterns from our 204-pattern library, we found that the Chi-square values associated with the 20,706 pairs of spectra are in all cases larger than 40. From this result, we reject the hypothesis of similarity between the 20,706 test pairs and infer that analytical error alone will not produce infrared absorption data sets that could be construed to be similar when in reality the data originate from different oils.

Figures 2 through 4 illustrate the degree of separation obtained with 20,706 possible pairs of spectra. The pairs shown are those for which 99.9%, 99%, and 95% of the 20,706 pairs produce larger Chi-square values. Thus Figure 2 illustrates the spectra corresponding to the 21st smallest χ^2 ($\chi^2 = 543.4$) and shows the spectra of Libyan and Julesburg, Colo., crude oils. Only 0.1% of all possible pairs exhibit closer similarity than the pair of spectra shown in Figure 2. Figure 3 is a similar illustration for the pair of spectra at the 99% level; i.e., only 1% of the 20,706 pairs are "closer" than Figure 3. The two oils represented by Figure 3 are Wesson field, Ark., and Fatem field, Dubai, crudes. Figure 4 shows the spectra of two used lubricating oils, with χ^2 values of 4,781.9.

Reducing Dimensionality. The measurement of all 21 variables may not be necessary for the purpose of identifying oils. There are undoubtedly some redundancies present in the

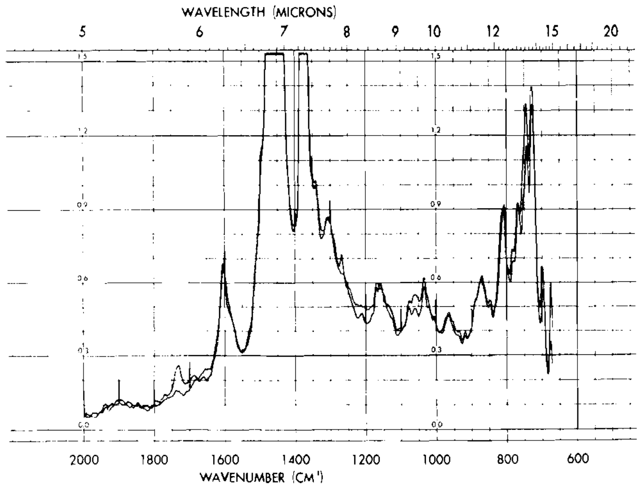

Figure 3. 99% of all possible pairs in the 204-oil library exhibit less similarity than these spectra of Wesson field, Ark., and Fatem field, Dubai, crude oils

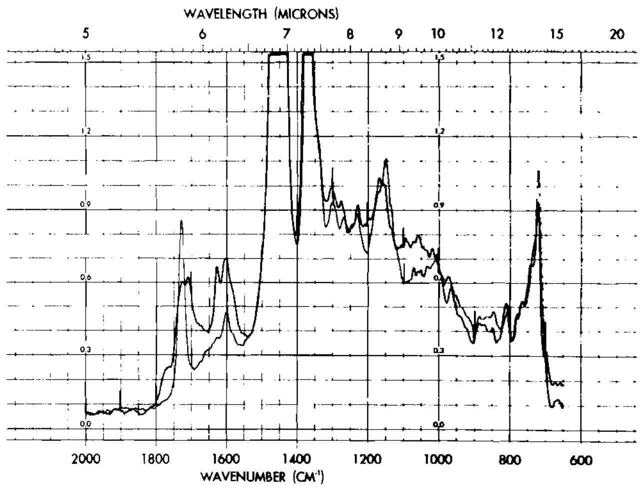

Figure 4. 95% of the pairs exhibit less similarity than do these two used crankcase lubricants

Table IV. Pearson Correlation Coefficients ≥0.90, Based upon 204 Oils, 21 Variables Each, Normalized to 0.1 mm Pathlength

Peak	1603	1166	1154	888	870	846	832	793
1166	...							
1154	...	0.99						
888	0.94	...	...					
870	0.95	...	...	0.98				
846	0.94	...	...	0.95	0.98			
832	0.93	...	...	0.96	0.99	0.99		
793	...	...	...	0.92	0.97	0.97	0.98	
781	0.91	...	...	...	...	0.91	0.90	0.94

Table V. Pearson Correlation Coefficients for Fourteen Variables Remaining after Elimination of All r ≥0.90

Peak	1629	1520	1304	1166	1032	963	918	832	809	765	741	722	697
1520	0.22												
1304	0.76	0.19											
1166	0.73	0.06	0.71										
1032	0.79	0.22	0.78	0.81									
963	0.63	0.09	0.65	0.86	0.81								
918	0.66	0.10	0.75	0.80	0.83	0.89							
832	0.54	0.30	0.56	0.24	0.70	0.35	0.47						
809	0.28	0.21	0.28	0.00	0.42	0.13	0.24	0.81					
765	0.32	0.23	0.28	0.06	0.46	0.10	0.20	0.72	0.64				
741	0.31	0.28	0.27	−0.02	0.46	0.09	0.19	0.84	0.83	0.84			
722	0.26	0.19	0.25	0.08	0.30	0.09	0.10	0.33	0.12	0.29	0.32		
697	0.09	0.23	−0.08	−0.13	0.22	−0.01	0.12	0.45	0.45	0.53	0.60	0.33	
672	0.09	0.17	0.05	0.04	0.08	0.05	0.07	0.00	0.16	0.11	0.01	0.50	0.25

infrared spectrum, considering the multitude of absorption bands produced by the average component molecule in a petroleum product or crude oil. In fact, Pearson correlation coefficients (*16*) computed for all 210 pairs of absorption bands for the entire 204-oil library reveal some high degrees of correlation between several peaks. Table IV lists those values of r, the correlation coefficient, which exceed 0.90. Excluding the anticipated high correlation between the 1166 and 1154 cm^{-1} peaks, all of the high r values involve the seven aromatic-related absorption bands at 1603, 888, 870, 846, 832, 793, and 781 cm^{-1}.

One of the requirements for employing the multivariate normal χ^2 test is that the variables be independent. Obviously, this is not the case for the 21-variable patterns described above. For example, a difference between two patterns in the 1165-cm^{-1} absorptivity should be reflected in a comparable difference in the 1154-cm^{-1} absorptivity, and including both measurements in a multivariate χ^2 test results in unintentionally weighting the information contained in either of those two peaks by a factor of two. One approach to solving this problem is to delete a sufficient number of absorption bands to eliminate high correlation coefficients. For an examination of Tables III and IV, one can use the values of s_p^2/s_a^2 to choose the more useful member of each highly correlated pair of variables. This results in the elimination of the 1154-cm^{-1} peak in favor of the 1166-cm^{-1} peak. The 793-cm^{-1} peak occurs as a distant maximum in only 5.4% of the spectra, and since it correlates highly with five other aromatic bands in the same region (888 to 781 cm^{-1}, average r = 0.96), it is not unreasonable to delete it from the pattern. Deleting the 781-cm^{-1} peak because of its low s_p^2/s_a^2 leaves the five highly correlated peaks (1603, 888, 870, 846, and 832 cm^{-1}), with an average r = 0.96. This problem can be resolved by retaining only one of the five variables in the final fingerprint. Because of its high s_p^2/s_a^2 (9,584), the 832-cm^{-1} peak is retained while those at 1603, 888, 870, and 846 cm^{-1} are deleted. The final pattern arrived at in the above procedure contains only fourteen variables, and there are no correlation coefficients among the 91 possible pairs ≥0.90. Table V lists the correlation coefficients for the final fourteen absorption bands. There are some relatively high values remaining among the fourteen final variables, but the average r in Table V is only 0.35. The deletion of the 1603, 1154, 888, 870, 846, 793, and 781-cm^{-1} peaks should have little effect on the ability to distinguish oils by the χ^2 test, as the average s_p^2/s_a^2 for the 14 variables is 1,985, compared to 1,863 for the original 21 variables.

Weathering. The preceding discussion has been restricted to fresh, unweathered oils with the exception of the 22 waste crankcase lubricants. Weathering of an oil spill at sea causes several changes to take place in the infrared spectrum of an oil. These include increases in many of the aromatic-related bands, the appearance of oxidation-related bands around 1700 cm^{-1} and disappearance of the 673 and 697-cm^{-1} bands. A major difficulty with weathered samples is the presence of water in the sample, which can often be removed by centrifugation at 30 °C and the addition of $MgSO_4$ (*14*). The χ^2 procedure described above is equally applicable to weathered oil spectra, by *considering weathering as a contributor to the estimated analytical variance,* s_a^2. It has been suggested (*14*) that a library of reference spectra might consist of spectra of oils which had been slightly artificially weathered, rather than fresh samples as used in this study. For actual field implementation of this procedure, the authors agree with that suggestion.

ACKNOWLEDGMENT

The authors acknowledge the helpful suggestions of Fredric Godshall of NOAA, Alan P. Bentz of the U.S. Coast Guard R & D Center, Morton Curtis of Rice University, and Chris Brown and Patricia Lynch of the University of Rhode Island.

LITERATURE CITED

(1) F. K. Kawahara, *Environ. Sci. Technol.*, **3**, 150 (1969).
(2) J. S. Mattson, *Anal. Chem.*, **43**, 1872 (1971).
(3) P. F. Lynch and C. W. Brown, *Environ. Sci. Technol.*, **7**, 1123 (1973).
(4) M. E. Garza and J. Muth, *Environ. Sci. Technol.*, **8**, 249 (1974).
(5) O. C. Zafiriou, *Anal. Chem.*, **45**, 952 (1973).
(6) D. E. Bryan, V. P. Guinn, R. P. Hackleman, and H. R. Lukens, "Development of Nuclear Analytical Techniques for Oil Slick Identification (Phase I)", Report GA-9889 Gulf General Atomic, Inc., USAEC Contract AT(04-3)-67 (1970).
(7) H. R. Lukens, D. Bryan, N. A. Hiatt, and H. L. Schlesinger, "Development of Nuclear Analytical Techniques for Oil Spill Identification (Phase IIA)", Gulf Radiation Technol., Report A10684, USAEC Contract AT(04-3)-67 (1971).
(8) M. Anbar, A. C. Scott, and M. E. Scolnick. "Identification of Oil Spills and Determination of Duration of Weathering by Field Ionization Mass Spectrometry", Abstract No. 224, Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, March 4–8, 1974, Cleveland, Ohio, 1974.
(9) A. D. Thurston and R. W. Knight, *Environ. Sci. Technol.*, **5**, 64 (1971).
(10) J. W. Frankenfeld, "Classification and Identification of Spilled Oil by Thin Layer Chromatography", Abstract No. 458, Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, March 3–7, 1975, Cleveland, Ohio, 1975.
(11) J. S. Mattson and C. A. Smith, "An On-Line Minicomputer System for Infrared Spectrometry", in "Computers in Chemistry and Instrumentation", Vol. 7, J. S. Mattson, H. B. Mark, Jr. and H. C. MacDonald, Jr., Ed., Marcel Dekker, New York, in press.
(12) A. Savitsky and M. J. E. Golay, *Anal. Chem.*, **36**, 1627 (1964).
(13) J. S. Mattson and A. C. McBride III, *Anal. Chem.*, **43**, 1139 (1971).
(14) C. W. Brown, University of Rhode Island, private communication, 1975.
(15) M. J. Spencer, "Oil Identification using Infrared Spectrometry", M.S. Thesis, University of Miami, School of Marine and Atmospheric Science, Miami, Fla, July 1975.
(16) W. C. Hamilton, "Statistics in Physical Science", Ronald Press, New York, 1964.
(17) D. F. Morrison, "Multivariate Statistical Methods", McGraw-Hill, New York, 1967.

RECEIVED for review January 9, 1976. Accepted October 7, 1976. This research was supported by the U.S. Coast Guard, Contract No. DOT-CG-81-75-1364 and DOT-CG-81-75-1383.