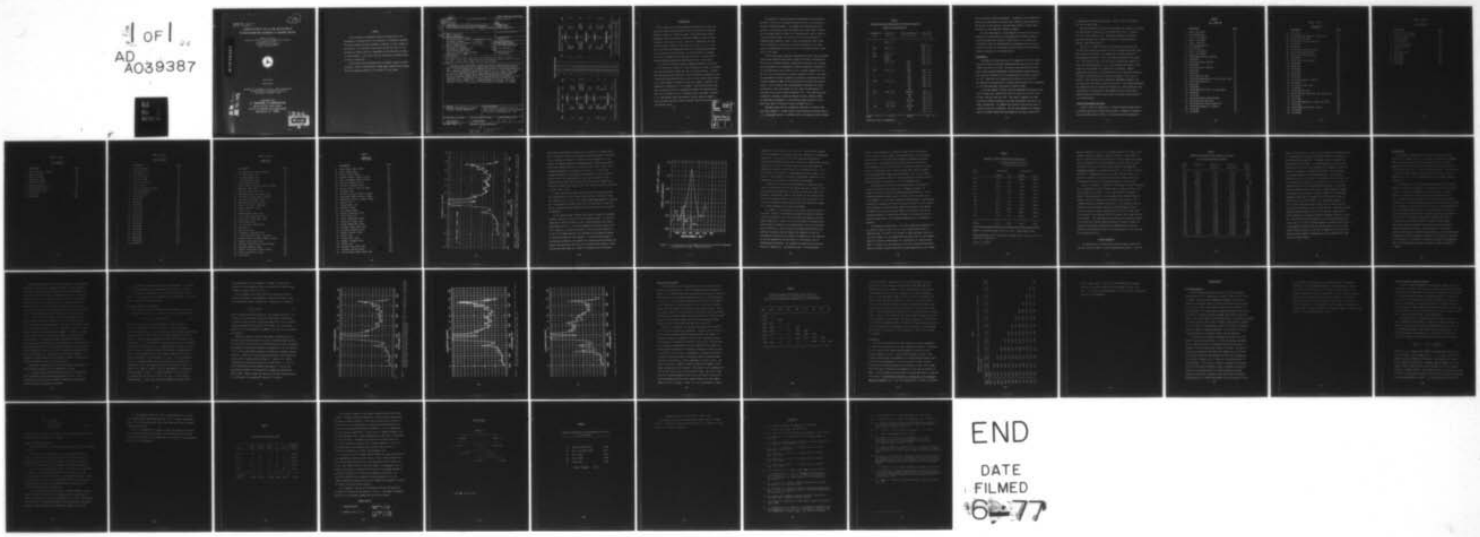
AD-A039 387 ROSENSTIEL SCHOOL OF MARINE AND ATMOSPHERIC SCIENCE --ETC F/G 11/8
CLASSIFICATION OF OILS BY THE APPLICATION OF PATTERN RECOGNITIO--ETC(U)
MAR 76 J S MATTSON DOT-CG-81-75-1364

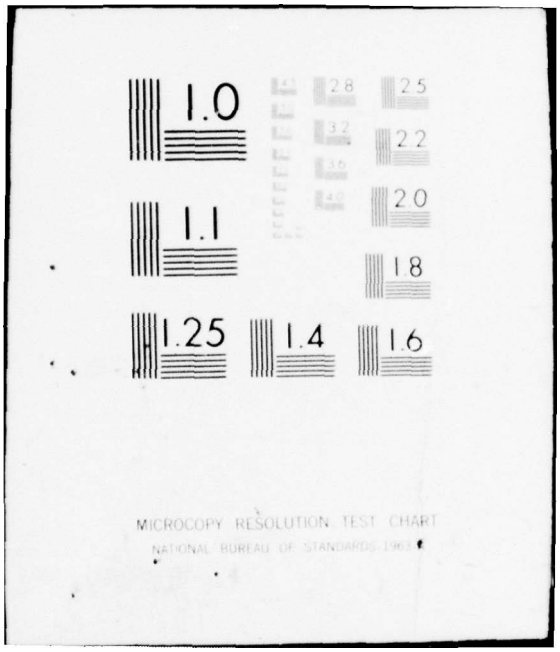
UNCLASSIFIED USCG-D-6-77 NL

1 OF 1
AD A039387



END

DATE FILMED
6-77


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963

(12)

REPORT NO. CG-D-6-77
Task No. 774243B.04

CLASSIFICATION OF OILS BY THE APPLICATION OF PATTERN RECOGNITION TECHNIQUES TO INFRARED SPECTRA

James S. Mattson
Rosenstiel School of Marine & Atmospheric Science
University of Miami
4600 Rickenbacker Causeway
Miami, Florida 33149

March 1976

Final Report

PREPARED FOR
U.S. DEPARTMENT OF TRANSPORTATION
UNITED STATES COAST GUARD
OFFICE OF RESEARCH AND DEVELOPMENT
WASHINGTON, DC 20590

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. The contents of this report reflect the views of the University of Miami, which is responsible for the facts and accuracy of data presented. This report does not constitute a standard, specification or regulation.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
USCG-D-6-77		

4. Title and Subtitle	5. Report Date
CLASSIFICATION OF OILS BY THE APPLICATION OF PATTERN RECOGNITION TECHNIQUES TO INFRARED SPECTRA.	March 1976
	6. Performing Organization Code
	8. Performing Organization Report No.

7. Author(s)	
James S. Mattson	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Rosenstiel School of Marine & Atmospheric Science University of Miami 4600 Rickenbacker Causeway Miami, Florida 33149	774243B.04
	11. Contract or Grant No. DOT-CG-81-75-1364
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	
Department of Transportation United States Coast Guard Office of Research and Development Washington, D. C. 20590	Final Report.
	14. Sponsoring Agency Code G-DOE-1/TP54

15. Supplementary Notes

The contract under which this report was submitted was under the technical supervision of the Coast Guard Research and Development Center, Groton, CT 06340. R&D Center report number CGR&DC 2/77 has been assigned.

16. Abstract

Parametric and nonparametric methods of pattern recognition have been used to group petroleum oils into six classes based solely upon infrared spectral data. A 204-oil training set, consisting of crude oils, fresh lubricating oils, waste crankcase lubricants, diesel fuel, and nos. 2, 4, 5, and 6 fuels, could be correctly classified in 82% of cases using a 3-nearest neighbor approach; 89% were correctly assigned using linear discriminant function analysis. Statistical analysis shows that the training set accurately represents the real population oils. A simplified χ^2 test is devised for the solution of the "infinite-class" problem; i.e., "fingerprinting" an oil spill. The logical extension of this χ^2 test would be its application to weathered oil samples.

17. Key Words	18. Distribution Statement
Infrared spectroscopy, oil classification, pattern recognition	This document is available to the U.S. public thru the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	49	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

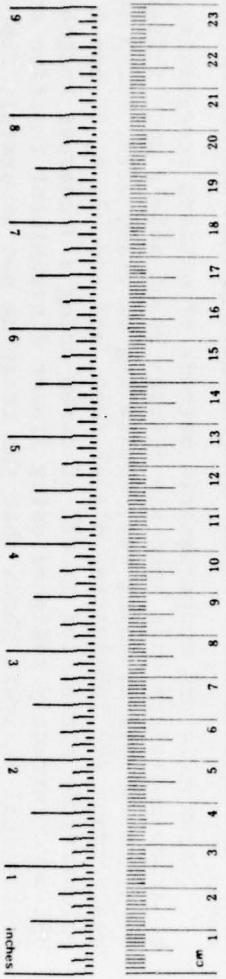
405 515

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

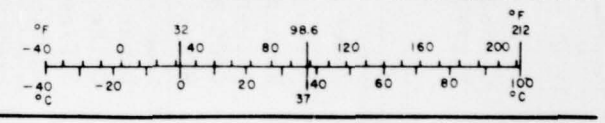
Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
	AREA			
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
	MASS (weight)			
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
	VOLUME			
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
	AREA			
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
	MASS (weight)			
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
	VOLUME			
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
	TEMPERATURE (exact)			
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F



INTRODUCTION

The classification of multicomponent petroleum oils (crude oils, lubricants, distillate and residual fuels) solely by their infrared absorption spectra is a difficult task. Crude oils alone include a phenomenal variety of systems, from heavy asphaltic crudes to light crudes that are similar to a No. 2 fuel oil. Furthermore, the distinctions between classes of fuel oils (i.e., Nos. 1, 2, 4, 5 and 6 fuels) are based upon ASTM specifications for continuous properties such as flash point and viscosity. In South Florida, for example, local fuel oil suppliers meet requests for Nos. 4 or 5 fuel oils by blending together appropriate proportions of Nos. 2 and 6 fuels.

In order to reduce the amount of sampling required in the event of an oil pollution incident, it would be useful to be able to initially classify the pollution sample into one of the above groups. Infrared spectroscopy has been promoted as a useful analytical technique for oil classification and identifications, since it does provide some information on the aliphatic, aromatic, polynuclear aromatic, carbonyl, and organosulfur composition of an oil (1-4). Infrared spectra have been used in previous efforts to distinguish asphalts from residual fuels (1), and to provide a tool for "fingerprinting" oils (2,3). Kawahara et al. (4) applied linear discriminant function analysis (LDFA) to their infrared data (1) to make the binary distinction between asphalts and residual fuels.

The problems of fingerprinting and classification are distinctly different, and will not necessarily involve the use of the same set of variables (absorption bands). For example, an absorption band that exhibits a low variance within each group, but a wide range of values over all groups taken as a whole (the "training set"), may be useful in a classification scheme, but could lead to an incorrect match in a fingerprinting situation. Table 1 lists the absorption bands considered by Kawahara (1), Mattson (2), and Lynch and Brown (3) in their previous investigations, as well as all of the absorption bands considered in this study.

In this study using primarily unweathered oils and a high-resolution, sophisticated computer-spectrometer system, the primary considerations were those of variable selection, experimental precision, and choosing between parametric and nonparametric pattern recognition techniques for the discrimination of six classes of oils. Pattern recognition is not new (5), but its application to chemical analysis is (6-10). Two powerful nonparametric pattern recognition techniques, learning machine and K-nearest neighbors (KNN), have been applied to both mass and infrared spectra. For a review of the learning machine technique, the reader is referred to Jurs and Isenhour (11,12), while the KNN approach is described in the recent review by Kowalski (13). The classical pattern recognition technique, linear discriminant function analysis (LDFA), is based upon an assumption of multivariate normal statistics within each class, and is discussed in detail elsewhere (14,15).

Efforts to apply pattern recognition techniques to chemical data have been centered to a large extent on major differences between groups; i.e., using mass spectral or infrared data to distinguish between carbonyl

TABLE 1

Infrared Absorption Bands Used for Pattern Recognition Studies of Petroleum Oils

	Kawahara (1) cm^{-1}	Mattson (2) $cm^{-1} \pm 1s_p$	Lynch and Brown (3) $cm^{-1} \pm 2.5\ cm^{-1}$	This Study $cm^{-1} \pm 1s_p$
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

and non-carbonyl-containing compounds. In addition, the original data have usually been derived from data files (Sadtler infrared spectra or API Project 44 mass spectra), and have been reduced to binary (peak/no-peak) or other simplified intensity formats.

For this investigation, it was apparent from the outset that the prediction of classes which are distinguished by different ranges of continuous properties (viscosity, flash point), not to mention the continuum of crude oil characteristics overlapping almost the entire range of fuel oils, is quite apart from the problem of identifying the presence or absence of a functional group.

EXPERIMENTAL

Transmission infrared spectra of 204 samples of oil were obtained using both precision sealed and demountable KBr cells 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 (16), 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}. The spectra were smoothed with a 21-point quartic smooth (17), normalized to a 0.10 mm pathlength, and stored on cassette tapes.

Of the 204 samples, 182 were fresh and unweathered lubricating oils (10), #2 fuels (30), diesel fuels (30), #4 fuels (12), #5 fuels (10), #6 fuels (28) and crude oils (62). The other 22 samples were waste automotive crankcase lubricants, obtained from service stations in the Miami, Florida area. All samples were stored at $5^{\circ}C$ from the time of receipt to the time of analysis. The 182 fresh samples were analyzed neat; the 22 waste lubes were centrifuged for one hour at about $35^{\circ}C$

to remove water and particulate matter. Table 2 lists the origins of the 204 oils where known.

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.

The use of a computer-interfaced Perkin-Elmer 180 assures that the analytical precision measured with this instrumentation would represent the best case possible for commercially available dispersive spectrophotometer systems. Thus, the results described in this report are a "best possible" case, and any effort to extend the results to a lower resolution instrument, or to a less sophisticated method of data reduction, will degrade the reliability of the method. Figure 1 graphically illustrates the reproducibility of the analytical technique. Six Kuwait crude replicates are plotted in Figure 1, one on top of another. These six spectra are completely independent analyses, not just six scans taken after filling the cell once. Only slight differences are apparent even to an experienced observer, but these differences are quantifiable and significant to the computer, and are particularly important when one considers the problem of matching infrared patterns in "fingerprinting".

VARIABLE AND BASELINE SELECTION

Table 1 lists the frequencies of infrared absorption bands used in previously reported oil identification studies. Kawahara (1) employed the seven bands listed in Table 1 in various nonlinear combinations,

TABLE 2

NO. 2 FUEL OIL

	Description	Oil #
1.	Amoco - 032 (USCG)	11
2.	Gulf Blended No. 2	13
3.	Conoco - 016 (USCG)	14
4.	Exxon - 057 (USCG)	15
5.	Shell - 010 USCG)	16
6.	Gulf Straight Run No. 2	17
7.	API "Pool" No. 2	18
8.	Texaco - 019 (USCG)	19
9.	Union - 005 (USCG)	20
10.	CH-012-F02 Baltimore, Maryland	73
11.	GU-080-F02	74
12.	ST-010-F02 Toledo, Refinery	75
13.	AM-035-F02	76
14.	SU-023-F02 Tulsa, Refinery	77
15.	BP-001-F02	78
16.	MO-014-F02	79
17.	ST-005-F02 Lima, Ohio	80
18.	EX-011-F02 Bayway Refinery, Linden, New Jersey	81
19.	SU-005-F02 Toledo Lab	82
20.	AM-013-F02	83
21.	MO-008-F02	84
22.	IN-095-F02CG City Coal Co., New London	85
23.	MA-003-F02	86
24.	CI-006-F02	87
25.	AR-022-F02 Houston, Texas Refinery	88
26.	TX-030-F02 Louisiana Plant	89
27.	HE-005-F02 Hess, Groton, Connecticut	90
28.	TE-003-F02 Chalmette, Louisiana	91
29.	GU-031-F02	92
30.	GE-004-F02	93

TABLE 2 (cont.)

DIESEL OILS

	Description	Oil #
31.	AR-013-D02	12
32.	IN-120-D02CG CGC Evergreen, New London	94
33.	ST-003-D02 Lima, Ohio	95
34.	EX-060-D02 O'Sullivan, New Haven	96
35.	EX-064-D02	97
36.	UN-013-D02	98
37.	CO-023-D02 Wirtlake, Louisiana	99
38.	PH-003-D02 Sweeny Refinery	100
39.	IN-017-D02CG CGC Point Knoll	101
40.	SH-038-D00	102
41.	ST-009-D02 Toledo, Refinery	103
42.	AM-016-D00	104
43.	SK-003-D00	105
44.	UN-002-D00	106
45.	EX-044-D00	107
46.	GE-007-D00	108
47.	MO-009-D00	109
48.	TX-032-D00 Lockport, Illinois	110
49.	GE-008-D00	111
50.	CH-010-D00 El Paw, Texas	112
51.	CI-008-D00	113
52.	ST-002-DPM Lima, Ohio	114
53.	EP-021-DOX Farm Bureau Ref., Mt. Vernon, Ind.	115
54.	EP-024-DOX	116
55.	TX-015-DOM	117
56.	EX-083-DOM Bayway Ref., Linden, New Jersey	118
57.	GU-075-MOD	119
58.	IN-003-MODSB USN Sub Base, Groton	120
59.	EX-006-MDF	121
60.	AS-009-MDF	122

TABLE 2 (cont.)

NO. 4 FUEL OIL

	Description	Oil #
1.	Amoco; Whiting, Indiana	21
2.	Chevron; El Paso, Texas	22
3.	Exxon-016 (USCG)	23
4.	Ind-016 (USCG)	24
5.	Sohio; Toledo, Ohio	25
6.	Sunoco-021 (USCG)	26
7.	Gulf #4	27
8.	Union-004 (USCG)	28
9.	CH-011-F04	123
10.	EX-079-F04	124
11.	EX-045-F04	125
12.	EX-059-F04	126

TABLE 2 (cont.)

NO. 5 FUEL OIL

	Description	Oil #
1.	Amoco Light No. 5 (USCG)	29
2.	Amoco-029 (USCG)	30
3.	Amoco-046 (USCG)	31
4.	Ashland-001 (USCG)	32
5.	Exxon-017 (USCG)	33
6.	Gulf Venezuelan No. 5	34
7.	Gulf Navy Special No. 5	35
8.	Sohio; Toledo, Ohio	36
9.	GU-032-F05	127
10.	GU-058-F05	128

TABLE 2 (cont.)

NO. 6 FUEL OIL

	Description	Oil #
1.	Amoco-025 (USCG)	37
2.	Gulf Blended No. 6	38
3.	API "Pool" No. 6	39
4.	Conoco-017 (USCG)	40
5.	Exxon-007 (USCG)	41
6.	Gulf Extra Heavy No. 6	42
7.	Ind-001 (USCG)	43
8.	Shell; Wood River, Illinois	44
9.	Gulf Straight Run No. 6	45
10.	Texaco-017 (USCG)	46
11.	IN-015-F06 DA	129
12.	EX-082-F06	130
13.	MA-004-F06	131
14.	AM-045-F06	132
15.	ST-012-F06	133
16.	CO-022-F06	134
17.	SH-045-F06	135
18.	BP-002-F06	136
19.	ST-007-F06	137
20.	GE-002-F06	138
21.	GU-034-F06	139
22.	GU-033-F06	140
23.	TX-025-F06	141
24.	CI-007-F06	143
25.	TX-026-F06	144
26.	MO-010-F06	145
27.	CH-007-F06	146
28.	AS-002-F06	147

TABLE 2 (cont.)

CRUDE OILS

	Description	Oil #
1.	Anchorage Basin, Alaska, Marathon	47
2.	S. Arabian Light, Arco	48
3.	S. Arabian Heavy, Gulf	49
4.	S. Arabian Medium, Gulf	50
5.	Cueta, Venezuela, Gulf	52
6.	Cotton Valley, Louisiana, Bureau of Mines	53
7.	Emeraude, Congo, Exxon	54
8.	Fairway Field, Texas, Shell	55
9.	Huntington Beach, California, Gulf	56
10.	Jay Field, Florida, Bureau of Mines	58
11.	Jay Smackover, Florida, Gulf	59
12.	Kern River, California, Shell	60
13.	Kuwait, API "Pool", Texas A&M	61
14.	Larosa, Venezuela, Exxon	62
15.	Libya, Gulf	63
16.	Opelika Field, Texas, Shell	64
17.	Orcutt Field, California, Shell	65
18.	Forrest Cy., Pennsylvania, EPA	66
19.	Pennsylvania, Gulf	67
20.	S. Tyler Field, Texas, Shell	68
21.	Utah, Gulf	69
22.	West Texas, Gulf	70
23.	Wilmington, California, Gulf	71
24.	S. Louisiana, API "Pool", Texas A&M	72
25.	Missouri, East Stutesbury, Bureau of Mines	148
26.	Alaska, North Slope, Conoco	149
27.	Arkansas, Wesson Field, Bureau of Mines	150
28.	Australia, Halibut, Amoco	151
29.	Montana, Bell Creek, Bureau of Mines	152
30.	Algeria, Zarzaitine, Conoco	153
31.	Peru, Union	154

TABLE 2

CRUDE OILS
(Continued)

	Description	Oil #
32.	Saudi Arabian Light, Exxon	155
33.	Dubai, Fatem, Exxon	156
34.	Sumatra, Minas, Texaco	157
35.	Wyoming, Recluse, Bureau of Mines	158
36.	Egypt, El Morgan, Bureau of Mines	159
37.	North Sea, Ekofisk, Conoco	160
38.	Colorado, Julesburg, Conoco	161
39.	Kansas, Bartlett, Bureau of Mines	162
40.	Trinidad, Amoco	163
41.	New Mexico, Justis, Bureau of Mines	164
42.	Mississippi, Moxie, Bureau of Mines	165
43.	Oklahoma, SW Enville, Bureau of Mines	166
44.	Nigeria, Gulf	167
45.	Iranian Heavy, Gulf	168
46.	Abu Dhabi, Exxon	169
47.	Iraq, Basrah, Exxon	170
48.	Alabama, Citronella, Gulf	171
49.	Libya, Brega, Gulf	172
50.	Alaska, Prudhoe Bay, Exxon	173
51.	Venezuela, W. Mara, Gulf	174
52.	Canada, Manyberries, Shell	175
53.	Michigan, Clare County, EPA	176
54.	Louisiana, Offshore, Gulf	177
55.	Ecuador, Oriente, Exxon	178
56.	Kuwait, Gulf	179
57.	Iranian Light, Shell	180
58.	Canada, S. Chavan, Shell	181
59.	Cabinda, Gulf	182
60.	Kentucky, Clay County, EPA	183
61.	Indonesia, Arojuna, Exxon	184
62.	Illinois Basin, Wayne County, EPA	185

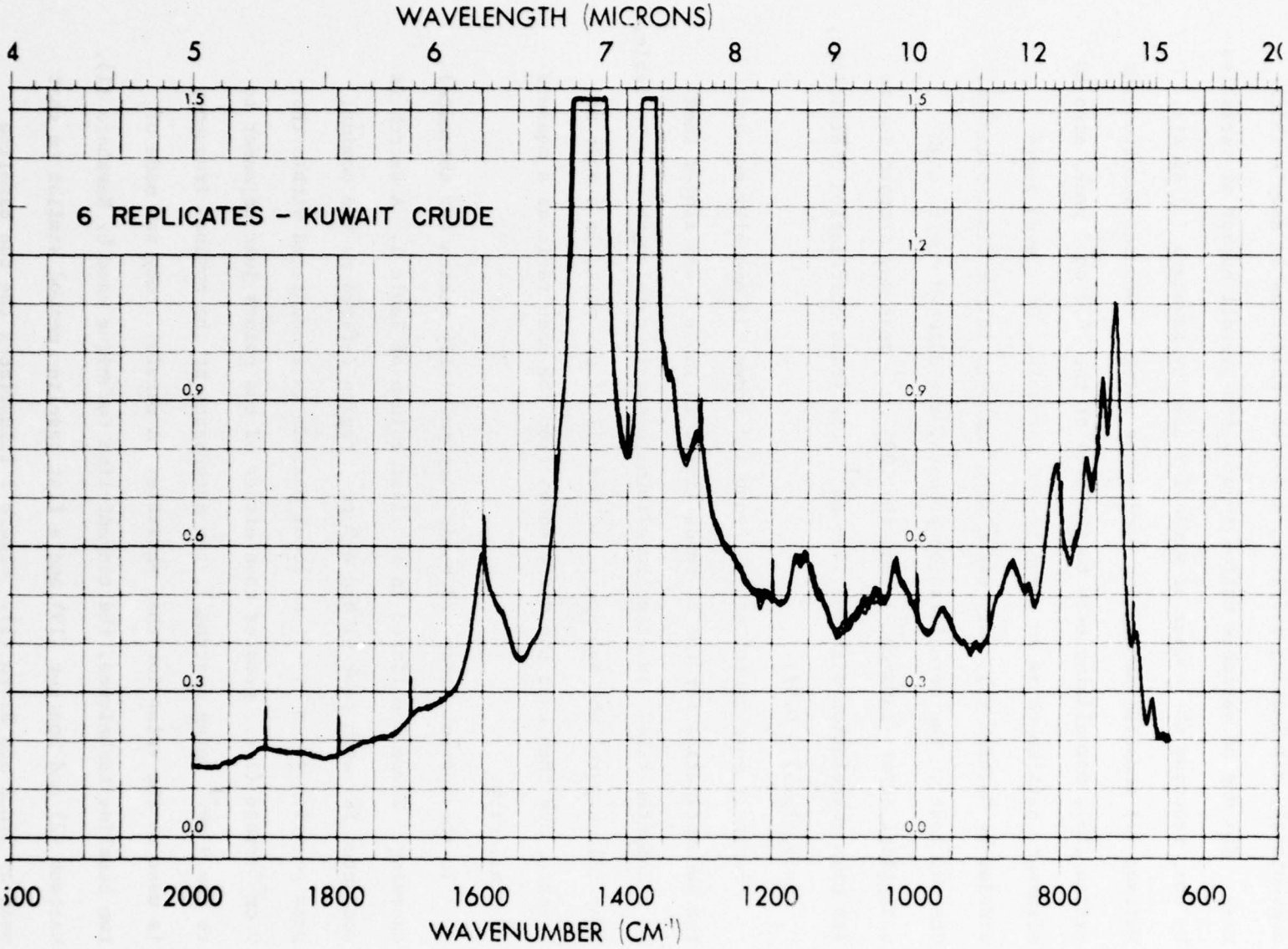


Figure 1. Computer-generated plots of six independent spectra of a Kuwait crude oil. The spectra have been smoothed once with a 21-point quartic smooth, resulting in a signal-to-noise ratio enhancement of about a factor of 2.

conveniently eliminating differences due to variations in sample thickness. The use of variable ratios reduces the overall number of variables by one, reducing the number of variables used by Kawahara (1) to six. Mattson (2) measured absorption band areas using a computer-spectrometer system (18), 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 1 and stored them in a computer file. In order to eliminate sample pathlength as a variable, they (3) computed the average ratio for all of the peaks in a pair of spectra, and then used the differences from the mean ratio as a measure of similarity.

In the instant study, 23 bands were initially chosen for automated encoding. These are listed in the last column of Table 1. A search is conducted for each peak within a 9 cm^{-1} region defined as the nominal peak position $\pm$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 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 (19) and a flat baseline method similar to that used by Lynch and Brown (3). Figure 2 illustrates the two baseline

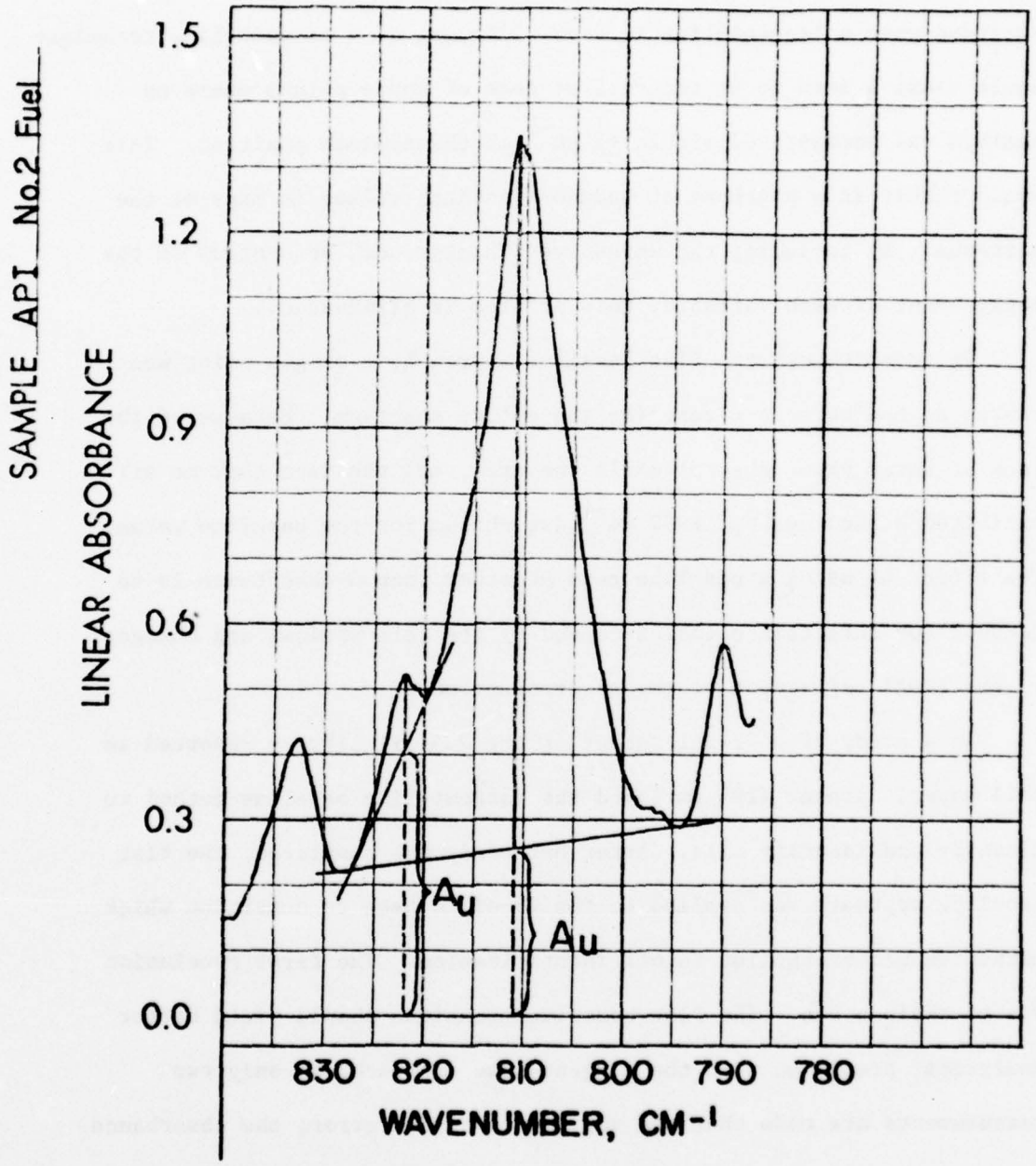


Figure 2. An illustration of the tangent-line and flat baseline approaches to measuring the height of absorption bands.

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 (19) employed the tangent-line baseline 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 units.

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 discernable to only the most experienced spectroscopists, sophisticated mathematical techniques often must be employed to make the final analysis. These mathematical methods 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 2, and ii) if so, can it improve the results in identification? In Table 3, 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.

TABLE 3

Comparison of Baseline Methods for Oil Spectra

Peak cm^{-1}	Estimated Standard Deviation, s in Absorbance Units x 10^3			
	Analytical[a]		Population[b]	
	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	14.7	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 $\equiv (\Sigma s^2)^{1/2}$

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 4 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 1. The column labelled "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 by analytical error.

TESTING NORMALITY

The prerequisite for employing multivariate normal statistics is that each variable exhibit a normal (Gaussian) distribution. Since we

TABLE 4

Frequency of Occurrence and Analysis of Variance

for 23 Infrared Absorption Bands

Peak cm^{-1}	Occurrence %	Relative s, analytical	Relative s, 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

intend to use these data for both oil classification and identification, it is important that the population density function be examined for any deviations from normality. The approach taken to evaluate the distribution was as follows. 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 population 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,..., 20) were computed (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 refined products.

FINGERPRINTING

The problem of identifying the source of an oil spill has been approached in a myriad of ways. At one time, it was seriously suggested that the conception 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 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 (20, 21), neutron activation analysis (22, 23), mass spectrometry (24), fluorescence spectrometry (25), and thin-layer chromatography (26). 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, could be amenable to statistical analysis using classical multivariate statistics or to more sophisticated, nonparametric pattern recognition techniques. This section is addressed to the problem of calculating as accurately as possible the probability that two infrared spectral patterns are similar enough to constitute a "match".

The envisioned application of the metric arrived at in the fashion described in this report is described as follows. Suppose there are ten possible suspects, any one of which may have been responsible for an oil spill, and that a sample is available from each of the suspects as well as the spill. If there were 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 spectra 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 section 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 of each variable. If the two are of the same magnitude, the usefulness of the variable is limited.

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

The Multivariate Normal Algorithm

For n dimensions, whether the probability distributions in any or all dimensions are normal or not, equation [1] defines a hyperellipsoid

$$\chi^2 = (\underline{X}_0 - \underline{X}_1)' \, \underline{M}^{-1} \, (\underline{X}_0 - \underline{X}_1) \qquad [1]$$

about the point $\underline{X}_0$ with a constant probability that is related to χ^2 (27). The matrix $\underline{M}^{-1}$ 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 is the standard deviation of the ith variable. Thus the diagonal elements of $\underline{M}$ 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 differences between the patterns is random analytical error in the measurement process, then several simplifications result. Since the differences in each variable Δx_1, Δx_2,..., Δx_n are assumed to be due to random measurement errors, the errors can be assumed to be independent, and the off-diagonal elements of $\underline{M}$ 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 4. A further assumption in such a case is to consider that the error vector $\Delta\underline{X} \equiv (\Delta x_1, \Delta x_2, \ldots, \Delta x_n)$ is multivariate normal. In that case, the metric computed with equation [1]

is distributed as X^2 with n degrees of freedom. This allows the calculation of the probability that two patterns are a "match" using ordinary univariate tables of X^2 (27).

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 [2]. Equation [2] is rigorously

$$X^2 = \sum_{i=1}^{n} (\Delta x_i / s_i)^2 \qquad [2]$$

valid only when the absolute values for the variables are known. For the 204 oils in the library, there are 20,706 possible unique pairs of patterns, which provides a sufficient sample for a test of possible erroneous matches due to analytical error alone. For 21 variables, values of $X^2 \geq 40$ correspond to a probability of 0.99 that two patterns are different.

Figures 3 through 5 illustrate the degree of matching obtained using all 20,706 possible pairs of spectra. The pairs shown are those for which 99.9%, 99%, and 95% of the 20,706 pairs show greater separation. Thus Figure 3 illustrates the spectra corresponding to the 21st smallest X^2 (X^2 = 543.4) and shows the spectra of Libyan and Julesburg, Colorado crude oils. Only 0.1% of all possible pairs exhibit closer similarity than the pair of spectra shown in Figure 3. Figure 4 is a similar illustration for the pair of spectra at the 99% level; i.e., only 1% of the 20,706 pairs are more similar than Figure 4. The two oils represented by Figure 4 are Wesson field, Arkansas and Fatem field, Dubai crudes. Figure 5 shows the spectra of two used lubricating oils, at the 95% point in increasing X^2 values (X^2 = 4,781.9).

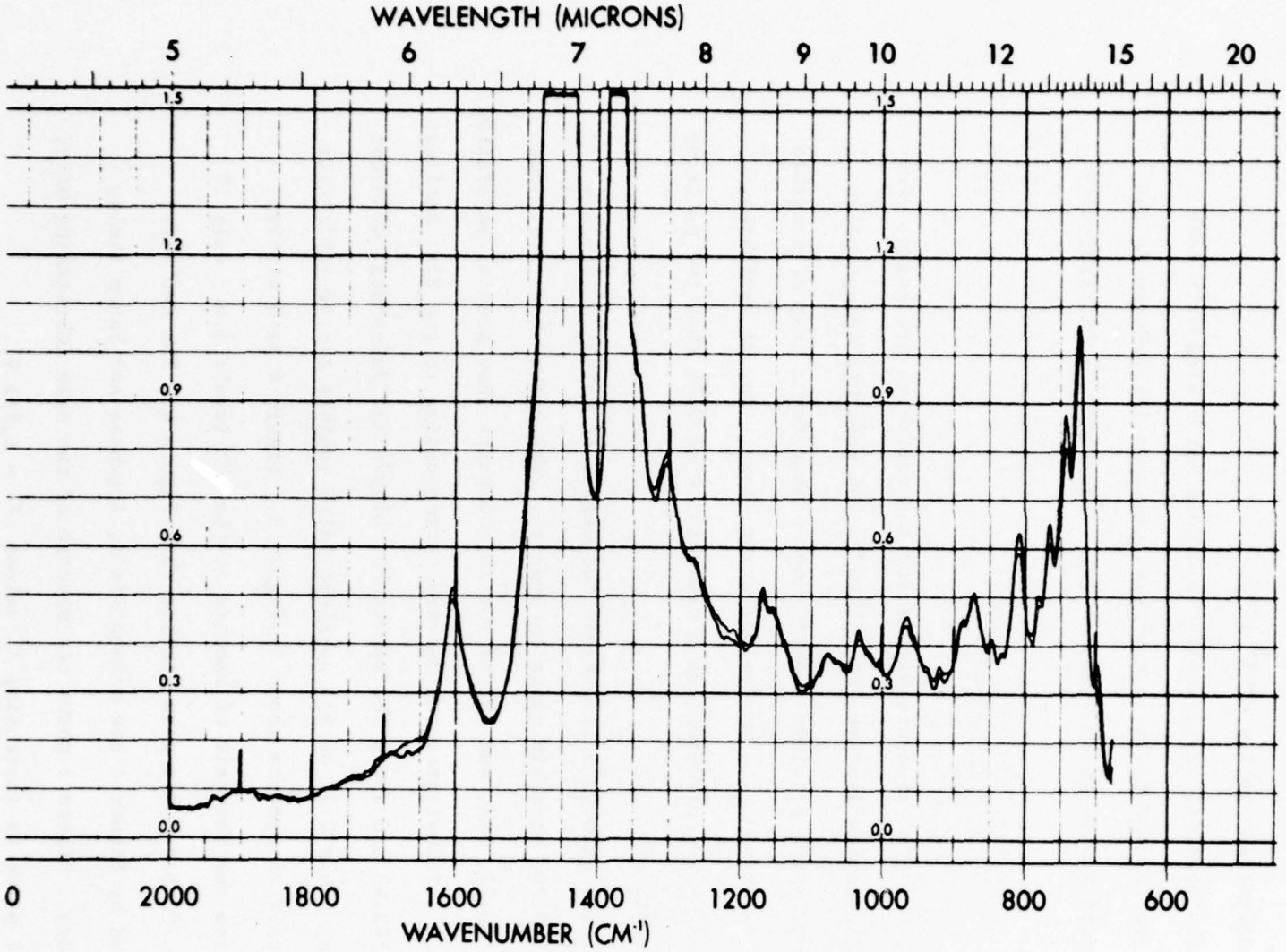


Figure 3. Infrared spectra of Libyan and Julesburg, Colorado crude oils, normalized to 0.1 mm pathlength, set equal at 1990 cm^{-1}, and 2X ordinate expansion. 99.9% of all possible pairs in the 204-oil laboratory exhibit less similarity than this pair.

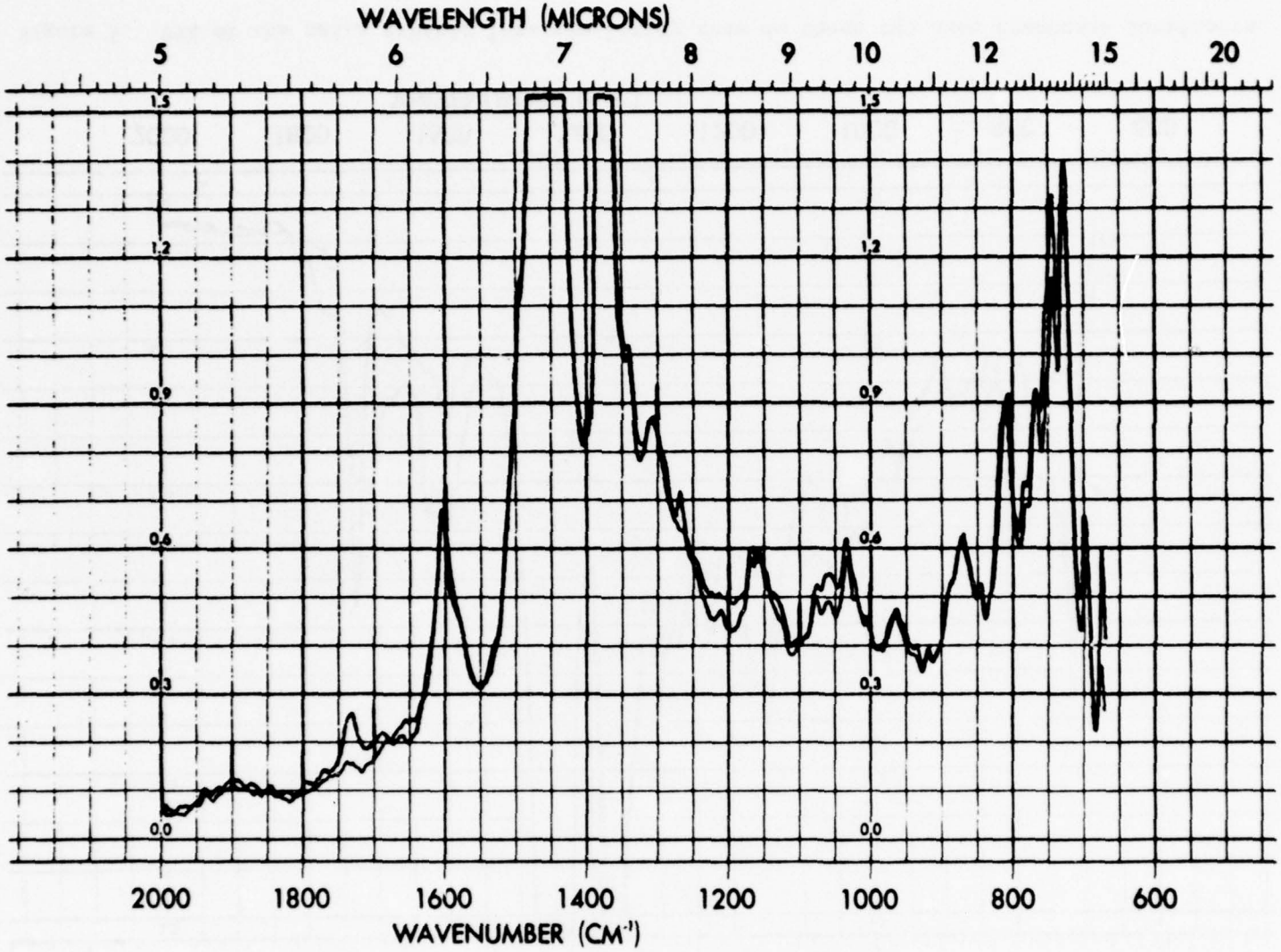


Figure 4. 99% of al possible pairs in the 204-oil library exhibit less similarity than these spectra of Wesson field, Arkansas and Fatem field, Dubai crude oils.

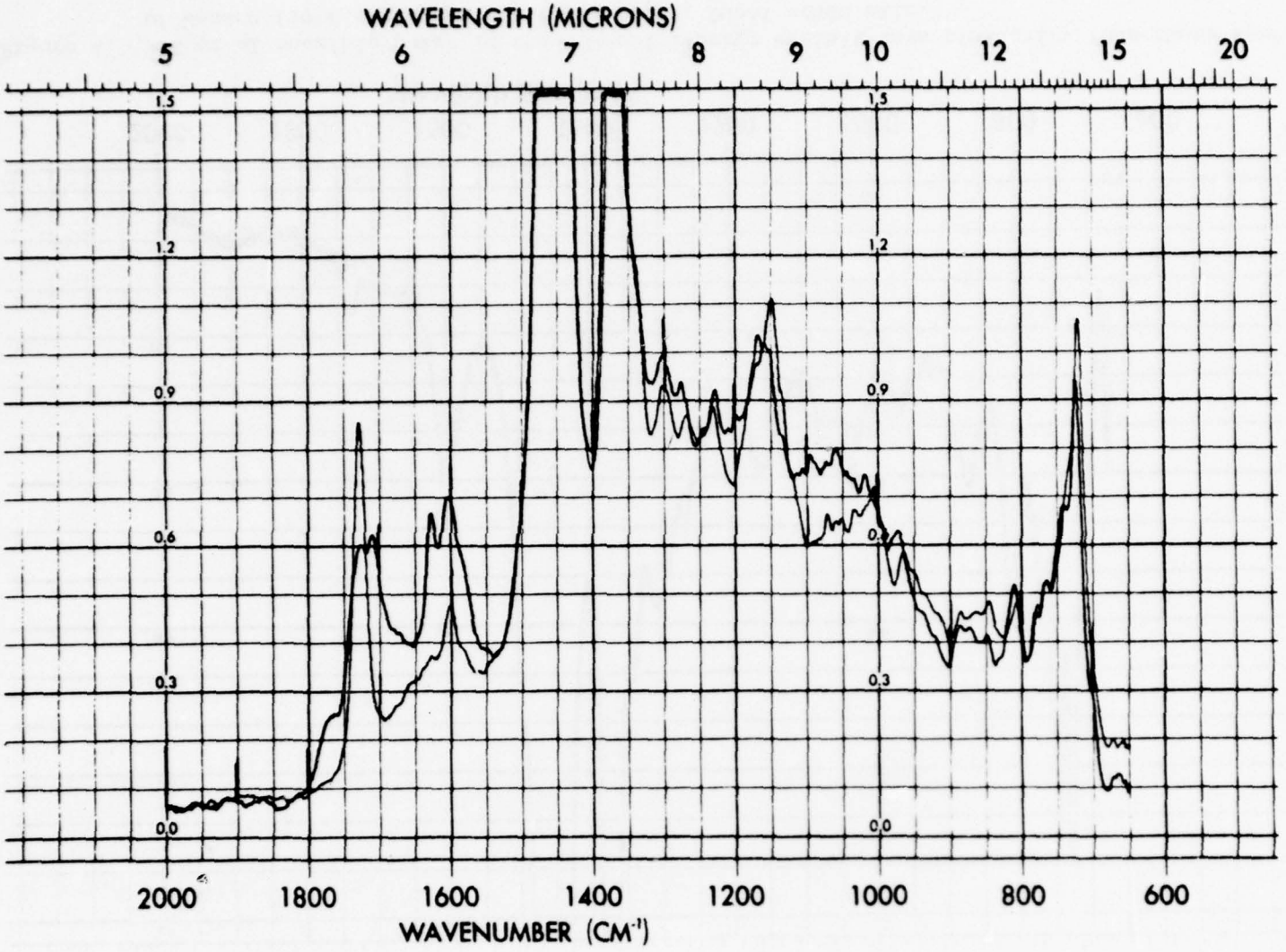


Figure 5. 95% of the pairs exhibit less similarity than do these two used crankcase lubricants.

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 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 (27) computed for all 210 pairs of absorption bands for the entire 204-oil library reveal some high degrees of correlation between several peaks. Table 5 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, 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. From an examination of Tables 4 and 5, 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

TABLE 5

Pearson correlation coefficients ≥ 0.90, based upon

204 oils, 20 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

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 $\geq$0.90. Table 6 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 6 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 X^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 $MgSO_4$. The X^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 that a library of reference

TABLE 6

Pearson correlation coefficients for fourteen variables remaining after elimination of all $r \geq 0.90$

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	-.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	-.08	-.13	0.22	-.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

spectra might consist of spectra of oils which had been slightly artifically weathered, rather than fresh samples as used in this study. For actual field implementation of this procedure, the authors agree with that suggestion.

CLASSIFICATION

Pattern Recognition

Pattern recognition involves the use of a series of similar observations made on a large group of objects, either with the intent to separate the objects into subgroups according to similarities between the individual objects, or with the intention of developing a set of rules by which future unknown objects can be classified into known subgroups. The former goal is called unsupervised learning, and can be used to identify previously unknown similarities between elements of the overall group. The principal method of unsupervised learning is cluster analysis, and is used to divide a large member of elements into two, three, or more subgroups, without any *a priori* knowledge of either the number of subgroups expected, or the property which most clearly separates the individual objects into the subgroups.

In this study, the number of subgroups were known, and a 204-member training set was available, in which the subgroup assignment of each element was known. The property which would separate the elements of the training set was, however, not known. The object of the pattern recognition study then is to examine all of the inter- and intra-group pattern relationships for the 204-oil training set, in an effort to develop a set of rules for the classification of future unknowns. This process, in which a training set consisting of objects with known properties is used to determine a predictive classifier for unknown objects, is called supervised learning. Testing the predictive classifier can be done using the straightforward approach of treating each member of the training set in turn

as an "unknown", and predicting its class based upon the information contained in the 203 remaining "knowns" in the training set. Kowalski (13) calls this the "leave-one-out" technique, and it is used in non-parametric approaches to pattern recognition. For the classical parametric method, linear discriminant function analysis (LDFA), the classifier functions are tested by their "recognition power" (6), or the ability to correctly classify all of the members of the training set. Kawahara et al. (4) used LDFA to classify asphalts and residual oils based on their infrared spectra, with a recognition power of >99%.

Linear Discriminant Function Analysis

Classical linear discriminant function analysis (LDFA) involves the derivation of classifier functions for each group, followed by predicting the probability of membership in each group. An unknown is then assigned to the group for which it shows the highest probability of membership. The validity of the classifier functions is tested by its "recognition power", or the ability to correctly classify the individual members of the training set. LDFA is based upon multivariate normal statistics, which we have shown to be the case for their data set, and it provides a unique set of classifier functions for a given training set.

The derivation of the classifier functions, which are linear in the n variables for each pattern, is based upon a weighting scheme which maximizes the contribution of those variables which are most effective in distinguishing a given class from the rest of the population. The classifier function for each group ℓ is in the form of equation [3], where Y_i^{ℓ} is related to the probability that the ith pattern belongs to

$$Y_i^{\ell}(x_1, x_2, \ldots x_n) = C_o + \sum_{j=1}^{n} W_j^{\ell} x_{ij} \qquad [3]$$

class ℓ, C_o is a constant term, and W_j^{ℓ} is the weighting coefficient for the jth variable. For m classes ($\ell = 1, 2, \ldots, m$), there are m sets of constant terms and weighting coefficients. The weighting coefficients W_j^{ℓ} are computed with the aid of a packaged computer program (14), which computes values of W_j^{ℓ} that favor those variables that show the greatest tendency to separate classes. The probability that a given pattern belongs to class ℓ is then given by equation [4], where Q_{ℓ} is the *prior probability* that an unknown belongs to class ℓ. All classes were given equal prior

$$P_i^{\ell} = \frac{Q_{\ell} \exp\ (Y_i^{\ell})}{\sum_{\ell=1}^{m} Q_{\ell} \exp\ (Y_i^{\ell})} \qquad [4]$$

probabilities in this test. An element is then classified along with the group for which the highest P_i^{ℓ} is obtained.

Results of LDFA Classification

A considerable amount of preliminary screening resulted in the following conclusion:

1. The 10 fresh lubricating oils were too indistinct and few in number to separate effectively, and the confusion they caused was too great when one considers that such materials do not pose a pollution problem. The fresh lube class was dropped from further consideration.

2. No. 4 fuels, another small group in this data set, are difficult to classify. The problem comes from the fact that some No. 4's are distillates and some are residual fuels. Some suppliers make "No. 4" fuel by blending No. 2 with No. 6. These problems made the No. 4 class one of the worst to deal with but it was eventually well classified by the "decision free" procedure described below.

3. No. 2 fuels and diesel fuels do not linearly separate. At best, using 10 variables, 50 out of the 60 samples could be classified into their correct groups, but the average probability of correct classification was only 0.563, not significantly greater than that expected from random guessing. These two groups were subsequently merged into one class.

4. The principal sources of errors in classification are i) crude oils classifying as lighter materials and, ii) No. 6 fuels classifying as crudes. All other groups (Waste lubes, Twos & Diesels, Fours, and Fives) classify without error.

Table 7 illustrates the "recognition power" and "predictive success" for a 5-group classification exercise. The "predictive success" along the bottom is indicative of the confidence one cal place in a classification prediction for an unknown oil.

TABLE 7

5-GROUP CLASSIFICATION BY LDFA

Group	Waste Lube	Two & Diesel	Four & Five	Six	Crude	Recognition Power
WLUBE	22	0	0	0	0	100.0%
Two & D	0	60	0	0	0	100.0%
4 & 5	0	0	22	0	0	100.0%
Six	0	0	2	25	1	89.3%
Crude	0	0	1	3	58	93.5%
Predictive Success	100.0%	100.0%	88.0%	89.3%	98.3%	96.4%

In an effort to improve on the 5-group classification results shown in Table 7 (187/194 correctly classified), a "decision tree" approach was developed, as shown in Figure 6. Any initial binary decision results in problems with such a wide variety of oils, but the resulting predictive success for the decision tree method is 189/194, or 97.4% an improvement over the 5-group classification. On the first cut (CRUDE/non-CRUDE), five oils are misclassed; 3 No. 6 fuels going with the crudes and 2 crudes going with the non-crudes. This makes the predictive confidence 98.5% for a "non-crude" prediction, and 95.2% for a "crude" prediction. All of the remaining stages of the decision tree classify without error; i.e., predictive confidences of 1.000 at each subsequent level.

For the 194 oils in 6 classes, Table 8 shows the average probability of class membership predicted for all oils of a class. These probabilities are limited by the values 0.985 for non-crudes and 0.952 for crudes, of course. What Table 8 shows is that even though it is possible to be, for example, 98.5% sure of a "waste lube" prediction, the average cumulative probability obtained for 22 actual waste lubes was only 89.8%. The cumulative probability is obtained by taking the product of all p's computed through the decision tree and the "predictive certainty" of either the crude or non-crude initial decision.

As an example of the use of this process, data were obtained from the USCG R & D Center for an unknown (to us) oil. This sample, designated #030-75 by Dr. Alan Bentz, passed down the tree as follows:

SAMPLE #030-75

1.	CRUDE/NON-CRUDE:	NON-CRUDE p = 0.944 CRUDE p = 0.056
2.	WLUBE/2 + D/4 + 5 + 6:	2 or Diesel p = 0.996 4, 5, or 6 p = 0.004 WLUBE p = 0.000

DECISION TREE

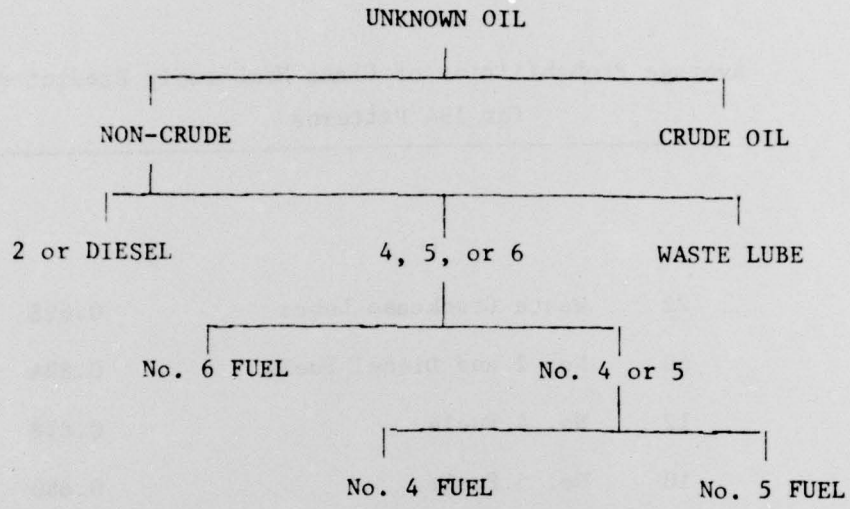


Figure 6. Decision Tree

TABLE 8

Average Probabilities of Class Membership Predicted for 194 Patterns

22	Waste Crankcase Lubes	0.898
60	No. 2 and Diesel Fuels	0.894
12	No. 4 Fuels	0.878
10	No. 5 Fuels	0.880
62	Crude Oils	0.838

Overall Average: 0.878

FINAL p (2 or D) = 0.996 x 0.944 x 0.985 = 0.926

Unfortunately, we have since learned that sample 030-75 is a light crude oil. However,the decision tree procedure is still expected to work for 95% of all crude oil samples.

REFERENCES

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. F. K. Kawahara, J. F. Santner and E. C. Julian, Anal. Chem., 46, 266 (1974).

5. G. S. Sebestyen, "Decision-Making Processes in Pattern Recognition", The Macmillan Co., New York, N. Y. 1962.

6. P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41, 690 (1969).

7. B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, ibid., 695 (1969)

8. B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, ibid., 1945 (1969).

9. P. J. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, ibid., 1949 (1969).

10. B. R. Kowalski and C. F. Bender, Anal. Chem., 44, 1405 (1972).

11. T. L. Isenhour and P. C. Jurs, in "Computers in Chemistry and Instrumentation", Vol. 1, Ed. by J. S. Mattson, H. B. Mark, Jr. and H. C. MacDonald, Jr., Marcel Dekker, Inc., New York, 1973, pp. 285-330.

12. P. C. Jurs and T. L. Isenhour, "Chemical Application of Pattern Recognition", Wiley, New York, 1975.

13. B. R. Kowalski, in "Computers in Chemical and Biochemical Research", Vol. 2, Ed. by C. E. Kloftenstein and C. L. Wilkins, Academic Press, New York, 1974, pp. 1-76.

14. W. J. Dixon, Ed., "Biomedical Computer Programs", University of California Press, Berkeley, 1974, pp. 221-254.

15. T. W. Anderson, "Introduction to Multivariate Statistical Analysis", Wiley, 1958.

16. J. S. Mattson and C. A. Smith, Ch. 2 in "Computers in Chemistry and Instrumentation", Vol 7, Ed. by J. S. Mattson, H. B. Mark, Jr. and H. C. MacDonald, Jr., Marcel Dekker, Inc., New York (in press).

17. A. Savitzky and M. J. E. Golay, Anal. Chem., 36, 1627 (1964).

18. J. S. Mattson and A. C. McBride III, Anal. Chem., 43, 1139 (1971).

19. M. J. Spencer, "Oil Identification using Infrared Spectrometry", M.S. Thesis, University of Miami, School of Marine and Atmospheric Science, Miami, Florida, July, 1975.

20. M. E. Garza and J. Muth, Environ. Sci. Technol., 8, 249 (1974).

21. O. C. Zafiriou, Anal. Chem., 43, 952 (1973).

22. 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).

23. 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).

24. 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 #224, 1974 Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, March 4-8, 1974, Cleveland, Ohio (1974).

25. A. D. Thurston, and R. W. Knight, Environ. Sci. Technol., 5, 64 (1971).

26. J. W. Frankenfeld, "Classification and Identification of Spilled Oil by Thin Layer Chromatography", Abstract #458, 1975 Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, March 3-7, 1975, Cleveland, Ohio (1975).

27. W. C. Hamilton, "Statistics in Physical Science", Ronald Press, New York, 1964.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1977--700-926--168