Analytica Chimica Acta, 112 (1979) 83—93
Computer Techniques and Optimization

INFORMATION THEORY APPLIED TO FEATURE SELECTION OF BINARY-CODED INFRARED SPECTRA FOR AUTOMATED INTERPRETATION BY RETRIEVAL OF REFERENCE DATA

P. F. DUPUIS**, P. CLEIJ, H. A. VAN 'T KLOOSTER* and A. DIJKSTRA

Analytisch Chemisch Laboratorium, Rijksuniversiteit Utrecht, Croesestraat 77A, Utrecht (The Netherlands)

(Received 4th July 1978)

SUMMARY

A method is described for feature selection from infrared spectra, intended for identification of organic compounds by computer-aided retrieval of reference data contained in small files. Complete discrimination of the binary-coded spectra is achieved by selecting a minimum number of spectral features; the information content is used as the selection criterion. The selection procedure is applied to five data sets (saturated and unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones) involving some 400 spectra. Each spectrum is uniquely coded by using about 10% of the 140 spectral features (binary-coded peak positions) available originally. For the intensity, a threshold of 50% appears to be applicable in some cases. For coding the frequency or wavelength parameter, wavenumbers (cm^{-1}) are preferred to wavelengths (μm). The method takes into account the a priori probabilities of spectral features and their correlations. Results of a retrieval program for a few "unknown" spectra are given.

Infrared spectrometry is an indispensable tool for identification of organic compounds. As in other areas of analytical chemistry, the vast numbers of spectra now available have emphasized the importance of computer-aided techniques, particularly for data processing and interpretation. Retrieval of reference spectra is one of the methods being used for automated identification [1].

In previous papers, the application of information theory to the coding of infrared spectra for retrieval purposes has been reported [2, 3]. Shannon's formula was used to calculate the information contents of retrieval procedures involving large files and different coding methods. It was shown that only a part of the spectral features contributes to the information content. It also appeared that errors occurring in the coded spectra and correlations between spectral features considerably diminish the information content and thus the applicability of infrared data collections, such as the (binary-coded) ASTM Infrared Spectral Index for retrieval purposes [2].

The application of numerical taxonomy and information theory to the selection of features from infrared spectra contained in files has recently been

**Present address: Dow Chemical (Nederland) B.V., Terneuzen, The Netherlands.

reported [4]. Numerical taxonomy was applied to the classification of peak positions with the correlation coefficient as a criterion of similarity. Features with relatively high calculated information contents were selected from groups of highly correlated peak positions. A file of 5100 spectra of various compounds, taken from the ASTM Index, was uniquely coded to the extent of 97.7% of the spectra, by selecting 40 out of 140 features (binary-coded peak positions); 99% of 395 spectra of hydrocarbons, alcohols, ethers and carbonyl compounds were uniquely coded by using 27 selected features.

This paper describes a method for unique coding of infrared spectra contained in small files with a minimum number of features; the information content is used as the selection criterion. The features are selected out of 140 possible peak positions, expressed in terms of the wavelength or the wavenumber, at intervals of 0.1 μm or 25 cm^{-1}, respectively. The method (called TREE) is applied to five data sets, each containing spectra of one type of compound (saturated and unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones). The underlying considerations on information theory are elucidated in the following paragraphs.

Uncertainty and information

Identification procedures involve gathering information, which can be considered as reducing uncertainty. Therefore, decrease of uncertainty with regard to the identity of the unknown compound (in this case, a pure organic compound) is used as a measure of the information obtained from the analysis. If it is assumed that, before analysis is carried out, there are n possibilities with equal probabilities for the identity of the unknown compound X, the uncertainty before analysis, $H(\mathrm{X})$, is, according to Shannon [5], given by:

$$H(\mathrm{X}) = \mathrm{ld}(n) \qquad (1)$$

with ld = $\log_2$. In the ideal case, the uncertainty will be reduced to zero and the information obtained will equal $H(\mathrm{X})$. Thus $H(\mathrm{X})$ can also be regarded as the "missing" or "required" information (for unambiguous identification).

In this study, infrared spectra are binary coded, thus for each feature F_j (peak position; j = 1—140) only two signals are possible: Y_{0j} and Y_{1j}, representing peak absent and peak present, respectively. When a signal Y_{k_j} is measured and provided that there are no errors, the uncertainty about the unknown identity will be reduced to a value

$$H(X/Y_{k_j}) = \mathrm{ld}(n_{k_j}) \qquad (2)$$

where n_{k_j} is the number of compounds giving a signal Y_{k_j} on analysis. The information gained is then defined by

$$I(X/Y_{k_j}) = -\Delta H(X/Y_{k_j}) = H(\mathrm{X}) - H(X/Y_{k_j}) \qquad (3)$$

Substitution of eqns. (1) and (2) into eqn. (3) gives

$$I(X/Y_{k_j}) = \mathrm{ld}(n) - \mathrm{ld}(n_{k_j}) = -\mathrm{ld}(n_{k_j}/n) \qquad (4)$$

The information content $I(F_j)$ of a feature (F_j) will be defined as the "expected value" of the information to be obtained:

$$I(F_j) = I(X/Y_j) = \sum_{k_j = 0,1} p(Y_{k_j}) \,.\, I(X/Y_{k_j}) \qquad (5)$$

where $p(Y_{k_j})$ is the probability of measuring a signal Y_{k_j}, when analysing the unknown compound. This probability can be calculated from

$$p(Y_{k_j}) = n_{k_j}/n \qquad (6)$$

Substitution of eqns. (4) and (6) into eqn. (5) yields

$$I(F_j) = -\sum_{k = 0,1} (n_{k_j}/n)\,\mathrm{ld}(n_{k_j}/n) \qquad (7)$$

If unknown and reference spectra are represented by series of features F_1, F_2,.....F_m, the information content of a retrieval procedure which searches and compares these coded spectra is given by

$$I(F_1, F_2, \ldots\ldots F_m) = -\sum_{k_1 = 0,1}\sum_{k_2 = 0,1}\cdots\cdots\sum_{k_m = 0,1} (n_{k_1, k_2, \ldots k_m}/n)\,\mathrm{ld}(n_{k_1, k_2, \ldots k_m}/n) \qquad (8)$$

where $n_{k_1, k_2, \ldots k_m}$ is the number of compounds in the reference file (= a priori possible identities) having a coded spectrum $(Y_{k_1}, Y_{k_2}, \ldots\ldots Y_{k_m})$. If the features (F_j) are not correlated, then eqns. (9) and (10) are valid:

$$n_{k_1, k_2, \ldots k_m}/n = (n_{k_1}/n)(n_{k_2}/n) \cdots\cdots (n_{k_m}/n) \qquad (9)$$

$$I(F_1, F_2, \ldots\ldots F_m) = \sum_{j=1}^{m} I(F_j) \qquad (10)$$

In most cases, however, correlations between features do occur and the information content becomes:

$$I(F_1, F_2, \ldots\ldots F_m) < \sum_{j=1}^{m} I(F_j) \qquad (11)$$

Accordingly, unique coding of all spectra in a file is possible only if the information content $I(F_1, F_2, \ldots\ldots F_m)$ equals the a priori uncertainty $H(X)$ (eqn. 1).

Feature selection by the TREE procedure

Efficient coding involves unique coding of the spectra representing the compounds of interest, with a minimum number of features and consequently a minimum use of computer bits for the storage of the spectra. In principle the optimal combination of peak positions can be determined by calculating the information contents for all possible combinations of m out of 140 peak positions. The number N of these combinations is given by $N = \Sigma_{m=1}^{140} \binom{140}{m}$. Although in practice iterative methods should greatly reduce this number, the remaining number of calculations will be too large to be handled by a computer. Therefore approximate methods must be used. Suitable approximations are provided by the selection procedure which we have called TREE.

In the TREE procedure, the information content of each of the 140 peak positions is calculated by using eqn. (7) and the one wih the highest information is selected. Next, the information contents of all 139 combinations of the first selected peak position with the remaining ones are calculated from eqn. (8). A second peak position is then selected, corresponding to the highest increment of the total amount of information. The third and following peak positions are selected in the same way. The procedure stops when the maximum increment of the information content equals zero, or when all spectra are uniquely coded. In this way correlations between peak positions are implicitly taken into account.

A flow chart of TREE is given in Fig. 1. The following example illustrates the TREE performance. Consider 6 compounds with their binary-coded spectra, each consisting of 5 features (see Table 1). From these data TREE selects 3 features as shown in Table 2.

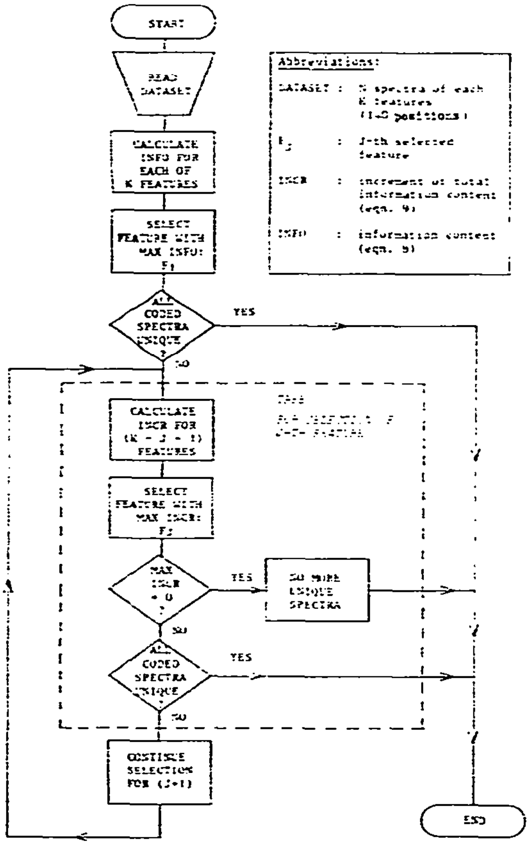
START
READ DATASET
CALCULATE INFO FOR EACH OF K FEATURES
SELECT FEATURE WITH MAX INFO: F1
ALL CODED SPECTRA UNIQUE ?
YES
NO
CALCULATE INCR FOR (K - J - 1) FEATURES
SELECT FEATURE WITH MAX INCR: FJ
MAX INCR = 0 ?
YES
NO MORE UNIQUE SPECTRA
NO
ALL CODED SPECTRA UNIQUE ?
YES
NO
CONTINUE SELECTION FOR (J+1)
END
Abbreviations:
DATASET : N spectra of each K features (140 positions)
FJ : J-th selected feature
INCR : increment of total information content (eqn. 9)
INFO : information content (eqn. 5)

Fig. 1. Flow chart of the TREE feature selection procedure.

TABLE 1

Binary-coded five-feature spectra of 6 compounds (hypothetical)

Compound	Feature				
	F_1	F_2	F_3	F_4	F_5
X_1	0	1	0	1	0
X_2	1	0	1	0	0
X_3	1	1	1	1	0
X_4	0	0	0	0	0
X_5	1	0	1	1	1
X_6	0	0	1	1	0

TABLE 2

Selection of features by TREE from data of Table 1

Step	Calculated information contents (eqns. 7 and 8) (bits)	Maximum information content (bits)	Maximum increment of total information content (bits)	Selected features	All spectra uniquely coded?
1	$I(1) = 1.00$ $I(2) = 0.92$ $I(3) = 0.92$ $I(4) = 0.92$ $I(5) = 0.65$	1.00	1.00	1	NO
2	$I(1, 2) = 1.92$ $I(1, 3) = 1.46$ $I(1, 4) = 1.92^a$ $I(1, 5) = 1.46$	1.92	0.92	1, 2	NO
3	$I(1, 2, 3) = 2.25$ $I(1, 2, 4) = 2.58$ $I(1, 2, 5) = 2.25$	2.58	0.66	1, 2, 4	YES
4	Procedure stops				

[a] In the algorithm used this value is discarded, as it does not exceed $I(1, 2)$.

EXPERIMENTAL

Data sets

The six data sets used are listed in Table 3. The infrared spectra were recorded and encoded by a standardized procedure, as described previously [3]. The spectral range was coded by using the wavenumber (WN code) and the wavelength (WL code); each resulted in 140 peak positions at intervals of 25 cm^{-1} and 0.1 μm, respectively. For the intensity threshold (IT), five different values (3, 5, 10, 25 and 50%) were used; this eventually resulted in 60 work files.

TABLE 3

Data sets used

Set	Number of spectra	Type of compounds
CHS	48	Saturated hydrocarbons
CHU	112	Unsaturated hydrocarbons
CH	160	Hydrocarbons (CHS + CHU)
ALC	100	Alcohols
ETH	66	Ethers
CARB	41	Aldehydes/ketones

Computer program

The computer program for TREE was written in Fortran IV and required about 60000 (octal) words of 60 bits. For testing and running, the CDC 73/26 computer of the Academic Computer Center of this University (ACCU) was used. The computer time T (in seconds) of a run depends mainly on the number (n) of spectra in the file and, with the sorting algorithm used, approximately satisfies the equation $T = 4.0 \times 10^{-3}\ n^2 + 0.25\ n$. The margin in T is about 15%, because of other variables. This equation was checked for n values smaller than 400. The second-order term is due to the presence of sorting procedures in the program, which are proportional to the square of the number of elements to be sorted. Thus the sorting procedure is the main limiting factor. Fast sorting routines, such as "heapsort" or "quicksort" [13], should extend the possibilities of the method to larger files. Although the computer time required for running the TREE program is substantial, it should be emphasized that in practice a feature selection procedure will be carried out only occasionally. The purpose is to provide not only a reliable but also a fast retrieval procedure, which is made possible by using a highly compressed code.

RESULTS AND DISCUSSION

The TREE procedure was applied to each of the 60 work files. An example of the results of the selection and coding procedures is illustrated in Fig. 2 for *trans*-4-octene. Out of the 140 binary-coded peak positions (WN code, IT = 10%, line b), TREE selected 12 features (line c), resulting in the reduced binary-coded spectrum given in line (d).

Table 4 shows the results of TREE for data set CHU (112 spectra of unsaturated hydrocarbons). When the WN code and an intensity threshold of 10% are used, the 112 spectra are all uniquely coded with 12 selected features. The total information content of the 12 selected features, calculated from eqn. (8), amounts to 6.81 bits, whereas addition of the information contents of the individual features, calculated from eqn. (7), yields 11.28 bits. The difference of 4.47 bits must be considered as due to correlation between the selected features. Table 5 shows the results obtained for all the data sets, WN- and WL-coded and with different intensity thresholds.

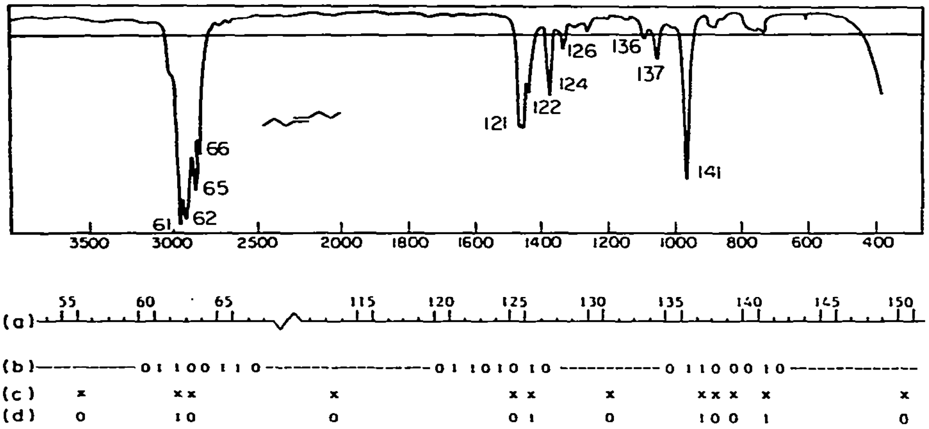

Fig. 2. The recorded infrared spectrum for *trans*-4-octene with the results of feature selection and coding procedures. Peak position numbers, corresponding to scale (a), are indicated on the spectrum. Line (b) shows the binary-coded peak positions (WN code) for the 10% intensity threshold indicated by the horizontal line on the spectrum; the dashes indicate zeros. Line (c) shows the features selected by TREE and line (d) the spectrum coded by TREE.

TABLE 4

Results of feature selection program TREE for the CHU data set (112 spectra of unsaturated hydrocarbons)

Information content $I(F_1....F_m)$ from eqn. (8) and number of uniquely coded spectra (n) for increasing number of selected features (m)			Information content $I(F_j)$ from eqn. (7), code and interval of spectral feature and sequence number of selected features (F_j)			
m	$I(F_1....F_m)$ (bits)	n	F_j	Code	Interval (cm^{-1})	$I(F_j)$ (bits)
1	1.00	0	1	63	2925—2901	1.00
2	1.99	0	2	125	1375—1351	0.99
3	2.94	0	3	139	1025—1001	0.99
4	3.87	0	4	138	1050—1026	0.99
5	4.77	6	5	56	3100—3076	0.94
6	5.49	15	6	141	975—951	0.92
7	6.05	42	7	131	1225—1201	0.92
8	6.40	70	8	137	1075—1051	0.97
9	6.61	90	9	150	750—726	0.90
10	6.72	102	10	126	1350—1326	0.95
11	6.77	108	11	62	2950—2926	1.00
12	6.81	112	12	113	1675—1651	0.71

Spectroscopic relevance of the selected features

It should be realized that the signals 1 (peak present) and 0 (peak absent) are equivalent in feature selection. At first sight, this chemometric approach differs from interpretation by an experienced spectroscopist, who looks

TABLE 5

Results of the feature selection program TREE. Percentages of uniquely coded reduced spectra (UCRS) and numbers of selected features (m) for all data sets, WN- and WL-coded with 5 different intensity thresholds.

Set	No. of spectra	Intensity threshold (%)	WN-code (cm^{-1})		WL-code (μm)	
			m	UCRS (%)	m	UCRS (%)
CHS	48	3	9	100	11	100
		5	9	96	11	96
		10	13	83	19	81
		25	9	50	9	23
		50	10	44	7	17
CHU	112	3	11	100	12	100
		5	11	100	13	100
		10	12	100	15	100
		25	15	100	22	96
		50	21	94	29	91
CH	160	3	12	100	16	100
		5	13	99	16	100
		10	17	95	25	93
		25	21	85	25	74
		50	23	78	33	66
ALC	100	3	10	100	13	100
		5	11	100	13	100
		10	10	100	12	100
		25	11	100	14	100
		50	13	100	17	100
ETH	66	3	9	100	10	100
		5	9	100	10	100
		10	10	100	10	100
		25	11	100	12	100
		50	11	94	13	94
CARB	41	3	8	100	9	100
		5	8	100	9	100
		10	8	100	9	100
		25	9	100	11	100
		50	12	100	14	95

primarily at the peaks present, though inferences are also made from the absence of peaks and spectra are interpreted on a retrieval basis from human memory as well as literature data. However, a distinction is usually made between features indicating functional groups and those in the fingerprint area (< 1400 cm^{-1}); such discrimination is not involved in the TREE procedure. Otherwise, it is noticeable that a considerable part of the selected features lies in the fingerprint area, as is illustrated by Table 4 for data set CHU.

Wavenumber and wavelength codes and intensity threshold

Data sets with WN-coded spectra require fewer features for unique coding than the corresponding WL-coded files (Table 5), because of the different numbers of peak positions per interval resulting from the use of the wavenumber or wavelength scale; particularly in the region around 3000 cm^{-1} more peak positions occur when the WN code is used.

The spectra of alcohols and aldehydes/ketones are all uniquely coded at an intensity threshold (IT) of 50% with the WN code. The complete data sets for unsaturated carbons (CHU) and ethers (ETH) are uniquely coded at an IT of 25%. The data sets for saturated hydrocarbons (CHS) and saturated and unsaturated hydrocarbons (CH) require an IT of 3% in order to obtain 100% uniquely coded spectra. These exceptional results are caused by the low number of intense peaks occurring in the spectra of saturated hydrocarbons.

Effects of errors in TREE-coded spectra on retrieval results

To illustrate the performance of the TREE procedure, a straightforward retrieval program was written to indicate the effect of errors in the coded spectra. The retrieval program was run on only one data set, viz. the file of 160 hydrocarbon spectra, which was considered to provide a fairly good illustration because the infrared spectra of hydrocarbons show considerable similarity. The number of "bit-mismatches" (mismatches of corresponding binary-coded features) between the TREE-coded "unknown" and the reference spectra was used as a match criterion; this was tested by using a logic XOR function. (The number of bit-mismatches can be considered as a measure of distance between two spectra). The maximum allowed number of bit-mismatches was taken as a variable, which was, at the start of the program, set to the number of selected features resulting from TREE.

Retrieval was done for eight infrared spectra which were not present in the reference data set (CH), but represented eight arbitrarily chosen compounds contained in the reference file. These "unknown" spectra were TREE-coded manually. Table 6 gives the results, showing the seven best matches for each "unknown" spectrum/compound. It appears that the correct answer is among the first 1—6 candidates, i.e. those with the lowest one or two numbers of bit-mismatches, with an average of 1.75. This indicates an error of about 5% in the coded unknown and/or reference spectra [2].

The first five candidates in each category indicate a certain classification power; work aimed at such classification is in progress. With respect to identification, the discriminating power of the procedure is diminished by minor errors. It should be stressed that this paper presents a method for the selection of the minimum number of spectral features, rather than an elaborate retrieval system of which feature selection is considered as a separate step. Therefore, none of the techniques described in the literature, e.g. windows [6] or weight factors [7], was employed; and extensive evaluation with a large test set of spectra based on criteria such as accuracy/precision/performance [8], recall/

TABLE 6

Results of retrieval program for 8 alternative spectra of reference compounds as unknowns. (CH data set (160 spectra of hydrocarbons); WN code; 10% intensity threshold; 17 selected features)

"Unknown"	7 best matches	No. of bit-mis-matches	"Unknown"	7 best matches	No. (bit-n matc
1-Heptene	1-Heptene	1	*o*-Xylene	Toluene	3
	2-Octene	2		*o*-Xylene	3
	1-Heptyne	2		1,2-Dimethylcyclohexane	4
	2-Methyl-2-pentene	2		Isopropylbenzene	4
	1-Hexene	2		*p*-Xylene	4
	n-Pentane	3		3-Methyl-2,4-pentadiene	4
	Cyclohexane	3		1-Phenylisobutene	4
n-Octane	n-Octane	0	1-Heptyne	1-Heptyne	0
	n-Hexadecane	0		1,3-Octadiyne	1
	n-Hexane	1		2-Methyl-2-pentene	2
	Cyclohexadecane	1		1-Hexene	2
	n-Decane	1		n-Pentane	3
	Methylcyclopentane	2		Cyclohexane	3
	1,2-Dimethylcyclohexane (c + t)	2		Methylcyclopentane	3
t-Butylbenzene	t-Butylbenzene	1	*m*-Xylene	1,3,5-Trimethylbenzene	2
	Methylcyclopentane	2		2,4,4-Trimethylpentane	3
	1,3,5-Trimethylbenzene	2		*m*-Xylene	3
	2,3,3-Trimethyl-1-butene	3		1,3-Dimethylcyclopentane (*c* + *t*)	3
	1,3-Dimethylcyclopentane (*c* + *t*)	3		1,3-Dimethylcyclopentane (*cis*)	3
	1,2-Dimethylcyclohexane	3		Pentaheptacontane	3
	1-Methyl-3-isopropylcyclopentane	3		Cyclohexadecane	4
Cyclohexane	Cyclohexane	2	Cyclohexadecane	1,3-Pentadiyne	3
	Decalin	2		Cyclohexadecane	4
	Benzene	2		Cyclooctane	4
	1,9-Decadiyne	2		t-1,3-Dimethylcyclohexane	4
	2,2-Dimethylhexane	3		1,3,5-Trimethylbenzene	4
	Cyclopentene	3		Diphenylmethane	4
	1-Pentyne	3		Adamantane	5

confusion [9], or recall/reliability [10, 11], was not considered to be relevant within the scope of this paper. The results of the straightforward retrieval program are presented mainly as an indication of the effect of errors in TREE-coded spectra on such results.

The effect of errors in binary-coded infrared spectra on information contents and retrieval results has already been reported [2]. An analogous study with binary-coded mass spectra has shown [12] that the efficiency of a retrieval system is determined far more by the extent to which errors occur in the (coded) spectra involved than by the matching criterion used, even if the latter takes some account of errors. There is no doubt that any effort at evaluation, optimization or development of retrieval systems must deal explicitly with these errors.

Conclusions

The information content is a useful criterion for feature selection of binary-coded infrared spectra. Unique coding of all the reference spectra in each of the data sets considered is possible by using about 10% of the available 140 peak positions. As the method takes into account the a priori probabilities of the peak positions and their correlations, the selection obtained is specific for the file considered.

The influence of the intensity threshold is similar for the data sets of unsaturated hydrocarbons, alcohols, ethers and aldehydes/ketones; values of up to 25% scarcely affect the number of selected features required for 100% uniquely coded spectra. An intensity threshold as high as 50% can safely be applied to the spectra of alcohols and aldehydes/ketones. However, there are considerable difficulties for the files containing saturated hydrocarbons, and only an intensity threshold of 3% gives meaningful results. All the data support the conclusion that use of the wavenumber scale is preferable to the wavelength scale in the coding procedure. A significant part of the selected features is situated in the fingerprint area.

The results indicate that automated identification by retrieval of spectra reduced and coded by the TREE procedure is successful if there are few errors in the coded unknown and/or reference spectra.

The authors are indebted to Dr. J. H. van der Maas, Mr. T. Visser and Mr. B. Lutz for kind supply of the data sets and valuable discussions.

REFERENCES

1 R. S. McDonald, Anal. Chem., 50 (1978) 282R.
2 P. F. Dupuis and A. Dijkstra, Fresenius Z. Anal. Chem., 290 (1978) 357.
3 P. F. Dupuis, J. H. van der Maas and A. Dijkstra, Fresenius Z. Anal. Chem., 291 (1978) 27.
4 F. H. Heite, P. F. Dupuis, H. A. van 't Klooster and A. Dijkstra, Anal. Chim. Acta, 103 (1978) 313.
5 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill., 1949.
6 Sirch III — Instruction Manual, American Society for Testing and Materials, 1969.
7 P. R. Naegeli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
8 D. S. Erley, Appl. Spectrosc., 25 (1971) 200.
9 S. L. Grotch, Anal. Chem., 45 (1973) 3.
10 G. M. Pesyna, R. Venkataraghavan, M. E. Dayringer and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
11 F. W. McLafferty, Anal. Chem., 49 (1977) 1443.
12 G. van Marlen, A. Dijkstra and H. A. van 't Klooster, Anal. Chem., in press.
13 D. E. Knuth, The Art of Computer Programming, Vol. 3. Sorting and Searching, Addison Wesley, Reading, Mass., 1973.