Journal of Chromatography, 294 (1984) 15–29
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 16,684

CHARACTERIZATION AND SELECTION OF STATIONARY PHASES FOR GAS–LIQUID CHROMATOGRAPHY BY PATTERN RECOGNITION METHODS*

J. F. K. HUBER* and G. REICH
Institute of Analytical Chemistry, University of Vienna, Waehringer Strasse 38, A-1090 Vienna (Austria)
(First received December 30th, 1983; revised manuscript received February 22nd, 1984)

SUMMARY

The classification of stationary phases in gas–liquid chromatography and the quantitation of their retention characteristics, which is generally described as 'polarity', were investigated by pattern recognition methods, especially by the hierarchical clustering and the minimum spanning tree techniques. It is demonstrated that the hierarchical clustering with a distant function and the minimum spanning tree method gives similar results with respect to the ranking and the differentiation of liquid stationary phases. New measures of the retention characteristics of liquid stationary phases were defined and tested. The potential of the various measures of solvent polarity is discussed. The mean retention index was found to be the best polarity characteristic. With a view to the rationalization of experimental work, an optimized procedure for the classification of liquid stationary phases and the calculation of their polarity with the minimum number of characteristic test solutes was elaborated. The set of characteristic solutes selected gives the best representation of the total number of solutes and covers all types of molecular interactions included in the term polarity.

INTRODUCTION

A nearly unlimited number of liquid stationary phases are available in gas chromatography if one considers that mixtures of stationary liquids can also be used. A classification of stationary phases and a drastic reduction in their number to a standard set are highly desirable in the interests of rationalization. The experimental choice of a single or several stationary phases for the separation of a given mixture by gas–liquid chromatography (GLC) should be confined to a search in the smallest possible number of solvents. In the identification of solutes by means of multi-dimensional GLC retention data, only stationary phases that make a significant contribution to the information content of the data should be applied. In order to achieve

* Presented in part at the Fourteenth International Symposium on Advances in Chromatography, Lausanne, September 24–28th, 1979.

0021-9673/84/$03.00

this objective, the number of liquid stationary phases should be reduced to those solvents which have significantly dissimilar retention characteristics. The systematic reduction of the number of stationary liquids requires their ranking in order of their effect on the retention pattern.

Several attempts to achieve this aim have been made. The first to try to characterize and classify stationary liquids for GLC was Rohrschneider[1,2] and his approach was refined and extended by McReynolds[3]. A few research groups have tried to quantify the similarity of stationary phases by means of pattern recognition methods. Leary and co-workers[4,5] calculated the Euclidian distances between phases and ranked them according to these values. Groups of phases are defined by their nearness and one member of each group is selected as representative of this group. This technique was also used by Haken *et al.*[6]. Massart and co-workers[7–10] used numerical taxonomy methods (cluster analysis and unsupervised learning methods are alternative names for this approach) to cluster the phases with a hierarchical cluster algorithm based on the correlation coefficient. Numerical taxonomy was also used by De Beer and Heyndrickx[11] to order the stationary phases with a selected set of solutes. Wold and co-workers[12,13] have used methods of principal components analysis (SIMCA) to calculate similarities between phases. Similar work was carried out by McCloskey and Hawkes[14]. Lowry *et al.*[15] used eigenvector projections to inspect the data set visually for spatial distribution in the pattern space. The data set used in all these papers, except refs. 9 and 11, was taken from McReynolds[3]. Factor analysis was used by Weiner and Parcher[16] and Dahlmann *et al.*[17] for the selection of preferred stationary phases. Haken and Srisukh[18] suggested a method for classifying stationary phases without the use of a reference stationary phase.

The aim of this work was to select the optimal pattern recognition method for the classification and selection of stationary phases in GLC. Two procedures, the hierarchical clustering method and the minimum spanning tree method, are included in the final evaluation. The criteria for selection of 'standard stationary phases' for GLC will be defined and a method for the selection of a limited number of test solutes will be developed. In all operations involved, only the retention value of the data base are used and no assumptions about the chemical structures of solutes or solvents are made. Several concepts for the characterization of solvent polarity by a single number are compared for the characterization of stationary phases in GLC. The polarity number suggested by McReynolds will be verified by a purely mathematical approach. A new concept, the mean retention index, will be shown to give equivalent results. Unfortunately, the data set used[19] on the one hand is based on a large number of obsolete stationary phases and on the other does not consider representatives of important types of compounds. It is, however, the only published data set that is sufficiently large and complete and other published data sets do not represent a complete data matrix without missing values. In order to obtain the full benefit of the proposed methods it would be necessary to create a new data base of relevant retention data. This point has also been stressed elsewhere[20].

EXPERIMENTAL

The calculations were carried out on a large computer (Control Data Cyber 170-720) with a main memory of 131K words of 60 bits and disk and tape storage.

The data and program input was performed by a remote process computer (PDP-15; Digital Equipment, Maynard, U.S.A.) with a core memory of 16K words of 18 bits, a dual magnetic tape drive (DEC-tape TU-56, DEC) and a line printer (Model 2200, Tally, Kent, U.S.A.). Both computers were connected by a dial-up telephone line via a modem (Model 300; Racal-Milgo, Reading, U.K.). Job editing, data entry and communication were performed on the PDP-15. The data bank was created and maintained on the Cyber. The main operating tool was ARTHUR, a program system for complex multi-dimensional data analysis by pattern recognition methods[21].

The data set used for the calculations was selected from a compilation given in the literature[19] containing relative retention values and retention indices of 367 solutes on 77 stationary phases at two temperatures. The greatest complete data matrix is formed by the retention data of 158 solutes on 75 stationary phases at 120°C. This set of data was obtained by exclusion of all solutes and stationary phases with missing values. The stationary phases of the selected data set are listed in Table I and the solutes in Table II.

RESULTS AND DISCUSSION

Classification procedure

In chromatography a solute is characterized by its retention values on n stationary phases, which means it represents a point in an n-dimensional space. The coordinates of this point form a so-called pattern vector X_i, which is defined by

$$X_i = x_{i1}, x_{i2}, \ldots, x_{in} \tag{1}$$

where the components x_{ip} are the retention indices of the solutes, i, on the stationary phases, $p = 1,2,\ldots,n$.

The stationary phases can also be characterized in this manner. A number of chemically different test solutes are used to characterize the retention characteristics of the stationary phase by the pattern vector X_p, which is defined by

$$X_p = x_{1p}, x_{2p}, \ldots, x_{kp} \tag{2}$$

where the components x_{ip} are the retention indices of the solutes, $i = 1, 2, \ldots, k$, on the stationary phase, p.

The stationary phases can be classified with respect to their retention characteristics by means of cluster analysis, a pattern recognition method. In this procedure the clustering of points in the k-dimensional space is investigated.

The algorithm of hierarchical clustering can be used for the classification of stationary liquids in GLC. The result of the classification by cluster analysis depends on the definition of the similarity measure. Two different similarity measures, the Euclidian distance and the correlation coefficient, were used. They are defined for the two solvents, p and r, as follows:

Euclidian distance:

$$d_{pr} = \sqrt{\sum_{i=1}^{k} (x_{ip} - x_{ir})^2} \tag{3}$$

TABLE I

LIST OF STATIONARY PHASES

No.	*Name*	*No.*	*Name*
1	Apiezon J	39	Neopentyl glycol succinate
2	Apiezon L	40	Oronite NIW
3	Apiezon M	41	Pluronic F68
4	Apiezon N	42	Pluronic F77
5	Bis(2-ethoxyethyl) phthalate	43	Pluronic F88
6	Carbowax 300	44	Pluronic L42
7	Carbowax 400	45	Pluronic L44
8	Carbowax 600	46	Pluronic L61
9	Carbowax 1000	47	Pluronic L63
10	Carbowax 1540	48	Pluronic L72
11	Carbowax 4000	49	Pluronic L81
12	Carbowax 6000	50	Pluronic P46
13	Carbowax 20M	51	Pluronic P65
14	Castorwax	52	Pluronic P84
15	Dibutyl tetrachlorophthalate	53	Pluronic P85
16	Diethylene glycol adipate	54	Polyphenyl ether, 5 rings
17	Diethylene glycol sebacate	55	Polyphenyl ether, 6 rings
18	Diethylene glycol succinate	56	Poly-tergent J-300
19	Di-2-ethylhexyl adipate	57	Quadrol
20	Di-2-ethylhexyl sebacate	58	SE-30
21	Diglycerol	59	SE-30 polyester NPGA terminated
22	Diisodecyl phthalate	60	SE-31
23	Dioctyl phthalate	61	SE-52
24	Dioctyl sebacate	62	Sorbitol
25	Dow Corning 550 fluid	63	Squalane
26	Dow Corning FS 1265 fluid	64	Sucrose acetate isobutyrate
27	Ethofat 60-25	65	Sucrose octaacetate
28	Ethylene glycol adipate	66	Tergitol NPX
29	Ethylene glycol sebacate	67	TMP tripelargonate
30	Flexol 8N8	68	Tricresyl phosphate
31	Hallcomid M18	69	Triethylene glycol succinate
32	Hallcomid M18 OL	70	Triton X-305
33	Hyprose SP 80	71	UCON LB-1715
34	Igepal CO 880	72	UCON 50 HB-2000
35	Isooctyl decyl adipate	73	Versilub F-50
36	Kroniflex THFP	74	XF 1150
37	Neopentyl glycol adipate	75	Zonyl E-7
38	Neopentyl glycol adipate terminated		

Correlation coefficient:

$$r_{pr} = \frac{\sum_{i=1}^{k} (x_{ip} - \bar{x}_p)(x_{ir} - \bar{x}_r)}{\sqrt{\sum_{i=1}^{k} (x_{ip} - \bar{x}_p)^2 (x_{ir} - x_r)^2}} \qquad (4)$$

TABLE II

LIST OF SOLUTES

No.	*Name*	*No.*	*Name*
1	Methanol	51	Isovaleraldehyde
2	Ethanol	52	2,2-Dimethylpropionaldehyde
3	Propanol	53	Hexanal
4	Isopropanol	54	Heptanal
5	Butanol	55	2-Ethylhexanal
6	Isobutanol	56	Acrolein
7	*sec.*-Butanol	57	Metacrolein
8	*tert.*-Butanol	58	Crotonaldehyde
9	Pentanol	59	2-Ethyl-2-butenal
10	Isopentanol	60	2-Ethyl-2-hexenal
11	2-Pentanol	61	Acetone
12	3-Pentanol	62	2-Butanone
13	2- Methyl-1-butanol	63	2-Pentanone
14	2-Methyl-2-butanol	64	3-Pentanone
15	3-Methyl-2-butanol	65	3-Hexanone
16	2,2-Dimethyl-1-propanol	66	3-Methyl-2-pentanone
17	Hexanol	67	4-Methyl-2-pentanone
18	2-Hexanol	68	3,3-Dimethyl-2-butanone
19	3-Hexanol	69	2-Heptanone
20	2-Methyl-1-pentanol	70	3-Heptanone
21	4-Methyl-1-pentanol	71	2-Octanone
22	2-Methyl-2-pentanol	72	Cyclopentanone
23	3-Methyl-2-pentanol	73	Cyclohexanone
24	4-Methyl-2-pentanol	74	3-Buten-2-one
25	2-Methyl-3-pentanol	75	5-Hexen-2-one
26	3-Methyl-3-pentanol	76	4-Methyl-3-penten-2-one
27	2-Ethyl-1-butanol	77	2,3-Butanedione
28	2,2-Dimethyl-1-butanol	78	2,4-Pentanedione
29	2,3-Dimethyl-2-butanol	79	Ethyl formate
30	3,3-Dimethyl-2-butanol	80	Propyl formate
31	3-Heptanol	81	Isopropyl formate
32	4-Heptanol	82	Isobutyl formate
33	2,2-Dimethyl-1-pentanol	83	*sec.*-Butyl formate
34	2,4-Dimethyl-3-pentanol	84	Pentyl formate
35	2-Octanol	85	2-Pentyl formate
36	2-Ethyl-1-hexanol	86	3-Pentyl formate
37	Cyclopentanol	87	Hexyl formate
38	Cyclohexanol	88	Methyl acetate
39	2-Propen-1-ol	89	Ethyl acetate
40	2-Propyn-1-ol	90	Propyl acetate
41	2-Buten-1-ol	91	Isopropyl acetate
42	3-Buten-2-ol	92	Butyl acetate
43	2-Methyl-2-propen-1-ol	93	Isobutyl acetate
44	1-Penten-3-ol	94	*sec.*-Butyl acetate
45	1-Penten-4-ol	95	*tert.*-Butyl acetate
46	Acetaldehyde	96	2-Pentyl acetate
47	Propionaldehyde	97	3-Pentyl acetate
48	Butyraldehyde	98	2-Methyl-2-butyl acetate
49	Isobutyraldehyde	99	Hexyl acetate
50	Valeraldehyde	100	4-Methyl-2-pentyl acetate

(*Continued on p. 20*)

TABLE II (*continued*)

No.	*Name*	*No.*	*Name*
101	2-Ethyl-1-butyl acetate	130	1,4-Butylene glycol formal
102	Heptyl acetate	131	Ethylene glycol acetal
103	Cyclohexyl acetate	132	1,3-Butylene glycol acetal
104	Allyl acetate	133	Diethyl propylal
105	Ethylene diacetate	134	Acrolein diethyl acetal
106	Methyl propionate	135	Pentyl ether
107	Propyl propionate	136	Tetrahydrofuran
108	Butyl propionate	137	2-Methyl-1,2-propylene oxide
109	Pentyl propionate	138	2-Methyltetrahydrofuran
110	Methyl butyrate	139	2-Methylfuran
111	Ethyl butyrate	140	2,5-Dimethyltetrahydrofuran
112	Isopropyl butyrate	141	Benzene
113	Butyl butyrate	142	Toluene
114	Pentyl butyrate	143	*o*-Xylene
115	Vinyl butyrate	144	*m*-Xylene
116	Butyl isobutyrate	145	*p*-Xylene
117	Isobutyl isobutyrate	146	Ethylbenzene
118	Methyl acrylate	147	*o*-Diethylbenzene
119	Ethyl acrylate	148	*m*-Diethylbenzene
120	Propyl acrylate	149	*p*-Diethylbenzene
121	Diethyl formal	150	Ethylene chloride
122	Isopropyl ethyl formal	151	Carbon tetrachloride
123	*sec.*-Butyl ethyl formal	152	Chloroform
124	Dibutyl formal	153	2-Chloroethanol
125	Ethylene glycol formal	154	3-Hydroxy-2-butanone
126	1,2-Propylene glycol formal	155	Dimethoxymethylal
127	1,3-Propylene glycol formal	156	1,4-Dioxane
128	1,3-Butylene glycol formal	157	Trioxane
129	2,3-Butylene glycol formal	158	1,3,5-Trioxepane

Both measures are used in the unweighted pair mode clustering algorithm.

The dendrogram of the unweighted pair mode clustering with the Euclidian distance as a similarity measure is shown in Fig. 1. The distance is used in the normalized form, s_{pr}, which is defined by

$$s_{pr} = 1 - d_{pr}/d_{pr\ \max} \tag{5}$$

where d_{pr} is the Euclidian distance from the pattern vector X_p to the pattern vector X_r and $d_{pr\ \max}$ is the maximum distance of two pattern vectors in the data set. This definition has the consequence that the similarity measure takes values between 0 (pattern with maximum distance) and 1 (two patterns at the same position). Two patterns are assumed to be most similar if their similarity value is larger than the values for any other pair of patterns. The dendrogram in Fig. 2 shows the result of the unweighted pair mode clustering with the correlation coefficients as the similarity measure. The assignment of the stationary phases to clusters can be done by splitting the dendrograms according to given similarity levels. The combining lines, which are cut by the chosen similarity level, delimit the clusters. The cluster numbers assigned to different stationary phases are shown in Table III. It can be seen that the clusters

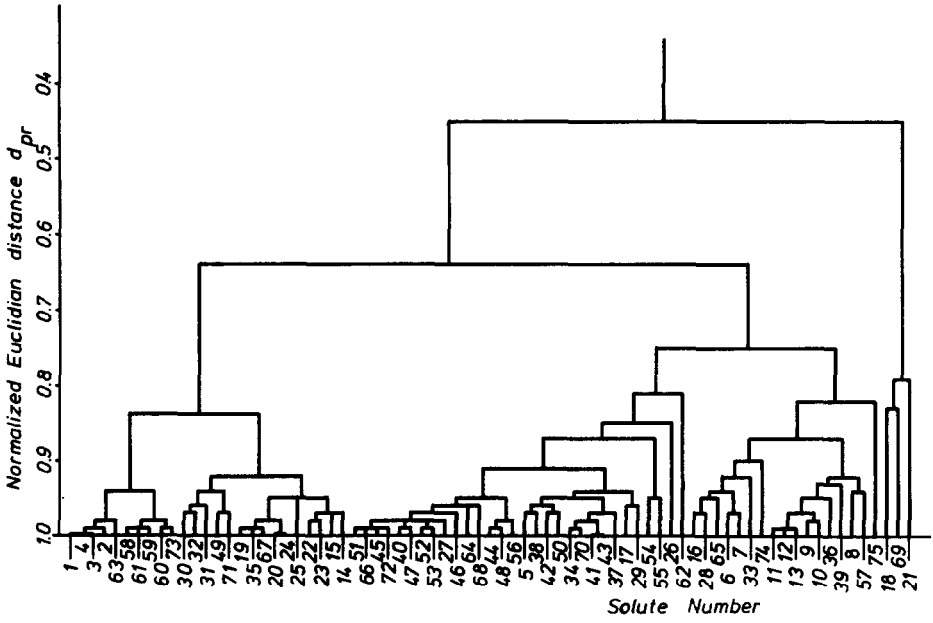

Fig. 1. Result of hierarchical clustering using the normalized Euclidian distance as a similarity measure.

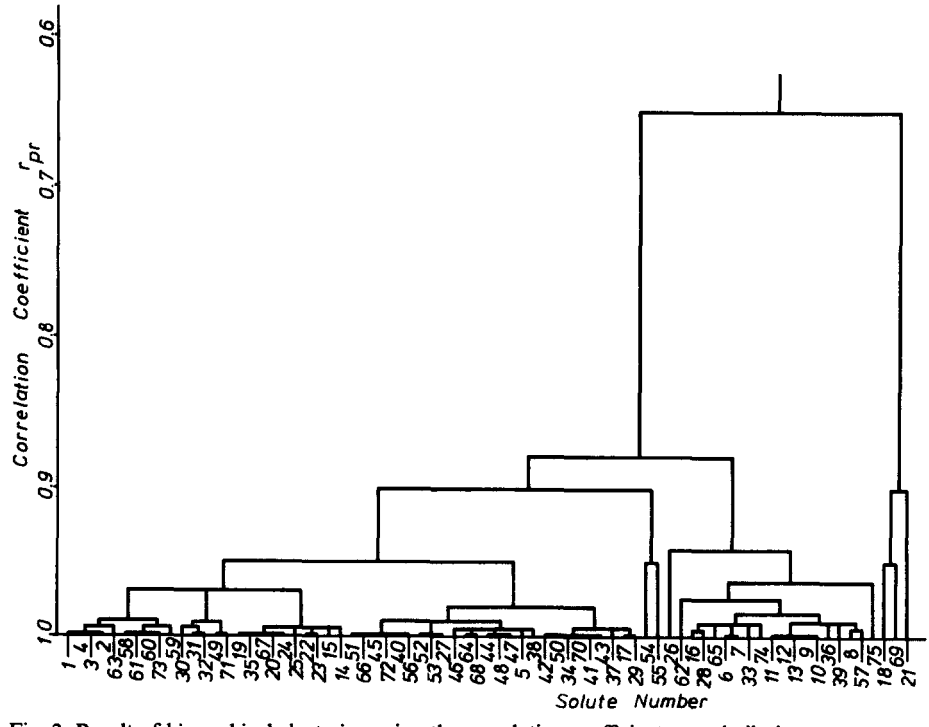

Fig. 2. Result of hierarchical clustering using the correlation coefficient as a similarity measure.

formed are identical for both similarity measures and there are only differences in the similarity levels.

Another method of cluster analysis that can be used is the minimum spanning tree method. In this method the pattern points are connected by a tree structure, for which the sum of the distances between consecutive points in the tree is a minimum,

TABLE III

RESULTS OF HIERARCHICAL CLUSTERING

Similarity values (s_{pr})				*Stationary phases forming the cluster*
0.90	*0.93*	*0.95*	*0.96*	
Cluster number				
1/2	1/2	1	1	Apiezon J, Apiezon L, Apiezon M, Apiezon N, squalane
		2	2	SE-30, SE-52, SE-30 polyester NPGA terminated, SE-31, Versilub F-50
3/7	3/5	3/5	3	Di-2-ethylhexyl adipate, isooctyl decyl adipate, TMP tripelargonate, di-2-ethylhexyl sebacate, dioctyl sebacate
			4	Diisodecyl phthalate, dioctyl phthalate, dibutyl tetrachlorophthalate, Castorwax
			5	Dow Corning 550 fluid
	6/7	6	6	Flexol 8N8, Hallcomid M18 OL, Hallcomid M18,
		7	7	Pluronic L81, UCON LB-1715
8/12	8/9	8/9	8	Pluronic L42, Pluronic L72, Poly-tergent J-300
			9	Pluronic P65, Tergitol NPX, Pluronic L44, UCON 50 HB-2000, Oronite NIW, Pluronic L63, Pluronic P84, Pluronic P85, Ethofat 60-25, Pluronic L61, sucrose acetate isobutyrate, tricresyl phosphate
	10/12	10/11	10	Bis(2-ethoxyethyl) phthalate, Neopentyl glycol adipate terminated, Pluronic F77, Pluronic P46
			11	Igepal CO 880, Triton X-305, Pluronic F68, Pluronic F88, neopentyl glycol adipate
		12	12	Diethylene glycol sebacate, ethylene glycol sebacate
13/14	13/14	13/14	13	Polyphenyl ether, 5 rings
			14	Polyphenyl ether, 6 rings
15	15	15	15	Dow Corning FS 1265 fluid
16	16	16	16	Sorbitol
17/21	17/19	17/18	17	Diethylene glycol adipate, ethylene glycol adipate
			18	Sucrose octaacetate
		19	19	Carbowax 300, Carbowax 400,
	20	20	20	Hyprose SP 80
	21	21	21	XF 1150
22/26	22/24	22/23	22	Carbowax 4000, Carbowax 6000, Carbowax 20M, Carbowax 1000, Carbowax 1540
			23	Kroniflex THFP
		24	24	Neopentyl glycol succinate
	25	25	25	Carbowax 600
	26	26	26	Quadrol
27	27	27	27	Zonyl E-7
28	28	28	28	Diethylene glycol succinate
29	29	29	29	Triethylene glycol succinate
30	30	30	30	Diglycerol

and closed loops in the connections are not allowed. This minimum spanning tree can then be split into clusters by an algorithm, which considers the distance of the adjacent points. This distance is defined as the Euclidian distance. There are three parameter values, which influence the pruning. These parameters describe the number of points included in the cutting decision, the limit for the variance of the distances and a normalization factor. Such a minimum spanning tree is shown in Fig. 3. The point positions of the stationary phases are calculated by the non-linear mapping method[22]. The points are projected from the 158-dimensional space to the 2-dimensional space, so that their inter-point distances represent reality as much as possible. The resulting clusters with the corresponding pruning parameters are given in Table IV.

Stationary phase selection

Up to this point only the classification of stationary phases into groups with similar retention characteristics has been performed. A simple decision is now to take from each cluster one phase that has favourable chemical properties, *e.g.*, high temperature stability, low viscosity and good reproducibility of the solvent characteristics, and can easily be obtained. It is then possible to define a standard set of stationary liquids that will solve most separation problems and give optimum data for the characterization of unknown solutes. An example of such a set of stationary phases that can be selected with this kind of procedure is shown in Table V. If for a given cluster several phases result as possible selections, then the phase with the minimum distance to the cluster centre should be chosen.

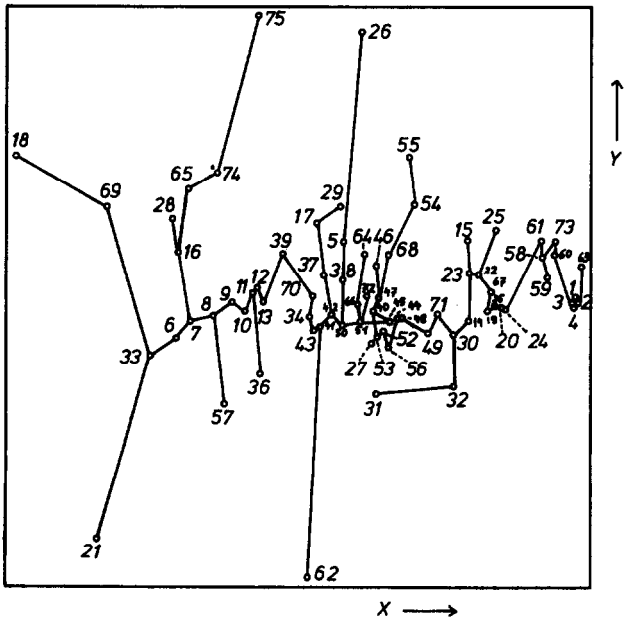

Fig. 3. Result of cluster analysis with the minimum spanning tree method projected to the two-dimensional space.

TABLE IV

COMPARISON OF CLUSTERING WITH DIFFERENT CLASSIFICATION METHODS

Stationary phase		*Classification method**				
No.	*Name*	*HC/0.90*	*HC/0.96*	*MST-4/1.5/0*	*MST-3/1.6/0*	*MST-3/1.2/0*
63	Squalane	1/2	1	1	1/2	1
2	Apiezon L	1/2	1	2	1/2	2
3	Apiezon M	1/2	1	2	1/2	2
4	Apiezon N	1/2	1	2	1/2	2
1	Apiezon J	1/2	1	2	1/2	2
60	SE-31	1/2	2	3	3	3
73	Versilub F-50	1/2	2	3	3	3
58	SE-30	1/2	2	3	3	3
59	SE-30 polyester NPGA terminated	1/2	2	3	3	3
61	SE-52	1/2	2	3	3	3
24	Dioctyl sebacate	3/7	3	4/5	4/5	4
20	Di-2-ethylhexyl sebacate	3/7	3	4/5	4/5	4
35	Isooctyl decyl adipate	3/7	3	4/5	4/5	4
19	Di-2-ethylhexyl adipate	3/7	3	4/5	4/5	4
67	TMP tripelargonate	3/7	3	4/5	4/5	4
22	Diisodecyl phthalate	3/7	4	4/5	4/5	5
25	Dow Corning 550 fluid	3/7	5	6	6	6
23	Dioctyl phthalate	3/7	4	4/5	4/5	5
15	Dibutyl tetrachlorophthalate	3/7	4	4/5	4/5	5
14	Castorwax	3/7	4	4/5	4/5	5
30	Flexol 8N8	3/7	6	4/5	4/5	5
32	Hallcomid M18 OL	3/7	6	4/5	4/5	5
31	Hallcomid M18	3/7	6	4/5	4/5	5
71	UCON LB-1715	3/7	7	4/5	4/5	5
49	Pluronic L81	3/7	7	4/5	4/5	5
48	Pluronic L72	8/12	8	7/10	7/10	7
44	Pluronic L42	8/12	8	7/10	7/10	7
56	Poly-tergent J-300	8/12	8	7/10	7/10	7
52	Pluronic P84	8/12	9	7/10	7/10	7
53	Pluronic P85	8/12	9	7/10	7/10	7
27	Ethofat 60-25	8/12	9	7/10	7/10	7
40	Oronite NIW	8/12	9	7/10	7/10	7
47	Pluronic L63	8/12	9	7/10	7/10	7
46	Pluronic L61	8/12	9	7/10	7/10	7
68	Tricresyl phosphate	8/12	9	7/10	7/10	8
54	Polyphenyl ether, 5 rings	13/14	13	11	11/12	11
55	Polyphenyl ether, 6 rings	13/14	14	12	11/12	12
45	Pluronic L44	8/12	9	7/10	7/10	7
51	Pluronic P65	8/12	9	7/10	7/10	7
72	UCON 50 HB-2000	8/12	9	7/10	7/10	7
66	Tergitol NPX	8/12	9	7/10	7/10	7
64	Sucrose acetate isobutyrate	8/12	9	7/10	7/10	9
50	Pluronic P46	8/12	10	7/10	7/10	10
38	Neopentyl glycol adipate terminated	8/12	10	7/10	7/10	10
5	Bis(2-ethoxyethyl) phthalate	8/12	10	7/10	7/10	10
26	Dow Corning FS 1265 fluid	15	15	13	13	13
42	Pluronic F77	8/12	10	7/10	7/10	10
37	Neopentyl glycol adipate	8/12	11	7/10	7/10	10

TABLE IV *(continued)*

Stationary phase		*Classification method**				
No.	*Name*	*HC/0.90*	*HC/0.96*	*MST-4/1.5/0*	*MST-3/1.6/0*	*MST-3/1.2/0*
17	Diethylene glycol sebacate	8/12	12	14	14/15	14
29	Ethylene glycol sebacate	8/12	12	15	14/15	15
43	Pluronic F88	8/12	11	7/10	7/10	10
41	Pluronic F68	8/12	11	7/10	7/10	10
62	Sorbitol	16	16	16	16	16
34	Igepal CO 880	8/12	11	7/10	7/10	10
70	Triton X-305	8/12	11	7/10	7/10	10
39	Neopentyl glycol succinate	22/26	24	17	17	17
13	Carbowax 20M	22/26	22	18/21	18/22	18
12	Carbowax 6000	22/26	22	18/21	18/22	18
11	Carbowax 4000	22/26	22	18/21	18/22	18
36	Kroniflex THFP	22/26	23	18/21	21	21
10	Carbowax 1540	22/26	22	18/21	18/22	18
9	Carbowax 1000	22/26	22	18/21	18/22	18
8	Carbowax 600	22/26	25	18/21	18/22	18
57	Quadrol	22/26	26	18/21	18/22	19
7	Carbowax 400	17/21	19	18/21	18/22	18
6	Carbowax 300	17/21	19	18/21	18/22	18
33	Hyprose SP 80	17/21	20	18/21	18/22	18
21	Diglycerol	30	30	24	24	24
69	Triethylene glycol succinate	29	29	22	18/22	22
18	Diethylene glycol succinate	28	28	25	25	25
16	Diethylene glycol adipate	17/21	17	18/21	18/22	18
28	Ethylene glycol adipate	17/21	17	18/21	18/22	18
65	Sucrose octaacetate	17/21	17	18/21	18/22	18
74	XF 1150	17/21	21	18/21	18/22	20
75	Zonyl E-7	27	27	23	23	23

* HC = Hierarchical clustering; MST = minimum spanning tree.

TABLE V

SELECTED STATIONARY PHASES

Stationary phase	*Selection procedure**				
	HC/0.90	*HC/0.96*	*MST-4/1.5/0*	*MST-3/1.6/0*	*MST-3/1.2/0*
	Cluster number				
Apiezon L	1/2	1	2	1/2	2
SE-30	1/2	2	3	3	3
Carbowax 400	17/21	19	18/21	18/22	18
Carbowax 20M	22/26	22	18/21	18/22	18
Diethylene glycol succinate	28	28	25	25	25
Polyphenyl ether, 6 rings	13/14	14	12	11/12	12
Tricresyl phosphate	8/12	9	7/10	7/10	8
Ucon 50 HB-2000	8/12	9	7/10	7/10	7
Di-2-ethylhexyl sebacate	3/7	3	4/5	4/5	4
Diisodecyl phthalate	3/7	4	4/5	4/5	5

* HC = Hierarchical clustering; MST = minimum spanning tree.

TABLE VI

RELATIONSHIP BETWEEN CLUSTERING AND MEASURE OF SOLVENT POLARITY

Stationary phase		*Cluster No.*	*Solvent polarity measure*			
No.	*Name*		*Euclidian distance*	*Distance in MST*	*Mean retention index*	*McReynolds constant*
63	Squalane	1	0	0	717.3	0
2	Apiezon L	1	1.17	1.17	720.1	42
3	Apiezon M	1	1.26	1.66	725.2	69
4	Apiezon N	1	1.37	1.89	726.6	74
1	Apiezon J	1	1.36	2.09	726.3	76
60	SE-31	2	2.53	4.18	743.1	109
73	Versilub F-50	2	2.90	4.71	746.4	126
58	SE-30	2	3.74	5.79	758.5	192
59	SE-30 polyester NPGA terminated	2	3.48	6.40	756.5	197
61	SE-52	2	4.08	6.38	761.3	200
24	Dioctyl sebacate	3	7.85	10.57	807.8	494
20	Di-2-ethylhexyl sebacate	3	7.86	10.70	808.0	498
35	Isooctyl decyl adipate	3	8.44	11.23	814.5	527
19	Di-2-ethylhexyl adipate	3	8.81	11.64	818.7	551
67	TMP tripelargonate	3	8.87	12.16	819.5	560
22	Diisodecyl phthalate	4	9.93	13.58	830.3	594
25	Dow Corning 550 fluid	5	8.64	15.75	811.0	470
23	Dioctyl phthalate	4	10.74	14.41	839.3	639
15	Dibutyl tetrachlorophthalate	4	11.06	15.58	841.3	673
14	Castorwax	4	10.98	15.98	844.3	716
30	Flexol 8N8	6	12.59	17.90	862.3	758
32	Hallcomid M18 OL	6	12.55	19.43	861.7	785
31	Hallcomid M18	6	10.97	21.20	842.6	691
71	UCON LB-1715	7	13.86	19.65	877.1	852
49	Pluronic L81	7	15.09	21.05	891.5	949
48	Pluronic L72	8	17.51	23.62	918.7	1080
44	Pluronic L42	8	17.97	24.15	924.3	1108
56	Poly-tergent J-300	8	18.36	25.14	929.0	1147
52	Pluronic P84	9	19.23	26.22	938.8	1202
53	Pluronic P85	9	19.88	26.95	946.2	1236
27	Ethofat 60-25	9	19.86	27.88	946.0	1238
40	Oronite NIW	9	19.88	27.82	945.8	1233
47	Pluronic L63	9	19.81	28.51	944.5	1197
46	Pluronic L61	9	20.22	29.59	947.0	1198
68	Tricresyl phosphate	9	19.00	30.32	933.2	1132
54	Polyphenyl ether, 5 rings	13	16.37	33.98	897.4	960
55	Polyphenyl ether, 6 rings	14	16.75	36.49	898.8	989
45	Pluronic L44	9	20.45	28.54	952.5	1262
51	Pluronic P65	9	20.89	29.12	957.5	1274
72	UCON 50 HB-2000	9	20.76	29.71	955.4	1269
66	Tergitol NPX	9	20.79	29.65	956.2	1290
64	Sucrose acetate isobutyrate	9	20.83	31.24	953.9	1250
50	Pluronic P46	10	22.66	30.94	977.9	1398
38	Neopentyl glycol adipate terminated	10	22.60	32.42	974.9	1398
5	Bis(2-ethoxyethyl) phthalate	10	22.84	33.86	975.6	1334
26	Dow Corning FS 1265 fluid	15	22.38	40.10	950.0	1110

TABLE VI *(continued)*

Stationary phase		*Cluster No.*	*Solvent polarity measure*			
No.	*Name*		*Euclidian distance*	*Distance in MST*	*Mean retention index*	*McReynolds constant*
42	Pluronic F77	10	24.03	32.40	993.1	1465
37	Neopentyl glycol adipate	11	24.70	33.94	999.0	1526
17	Diethylene glycol sebacate	12	25.36	36.38	1006	1571
29	Ethylene glycol sebacate	12	23.29	38.59	982.2	1444
43	Pluronic F88	11	25.31	33.90	1008	1573
41	Pluronic F68	11	25.27	34.08	1007	1571
62	Sorbitol	16	27.60	42.06	1031	1925
34	Igepal CO 880	11	25.91	34.71	1014	1597
70	Triton X-305	11	26.17	35.28	1017	1616
39	Neopentyl glycol succinate	24	28.69	38.28	1044	1756
13	Carbowax 20M	22	30.70	40.94	1069	1893
12	Carbowax 6000	22	30.95	41.54	1071	1907
11	Carbowax 4000	22	31.37	42.03	1076	1939
36	Kroniflex THFP	23	30.96	44.30	1073	1921
10	Carbowax 1540	22	32.23	43.08	1087	1996
9	Carbowax 1000	22	33.23	44.14	1098	2058
8	Carbowax 600	25	34.80	45.88	1117	2177
57	Quadrol	26	34.26	48.88	1109	2122
7	Carbowax 400	19	37.41	48.57	1146	2325
6	Carbowax 300	19	38.71	49.94	1161	2419
33	Hyprose SP 80	20	41.19	53.03	1189	2599
21	Diglycerol	30	46.37	61.95	1249	3111
69	Triethylene glycol succinate	29	45.65	59.36	1234	2818
18	Diethylene glycol succinate	28	54.17	68.36	1324	3261
16	Diethylene glycol adipate	17	38.42	51.40	1153	2372
28	Ethylene glycol adipate	17	38.63	52.80	1154	2364
65	Sucrose octaacetate	18	37.96	53.83	1146	2265
74	XF 1150	21	35.04	57.75	1112	2094
75	Zonyl E-7	27	31.62	64.60	1055	1673

TABLE VII

QUALITY OF LINEAR REGRESSION

Polarity measure	*For 10 selected compounds*		*For maximum number of selected compounds*		
	Remaining variance (%)	*Correlation coefficient*	*Max. number of compounds selected*	*Remaining variance (%)*	*Correlation coefficient*
McReynolds constant (*MR*)	0.187	1.0000	29	0.082	1.0000
Euclidian distance (*ED*)	1.180	0.9997	26	0.654	0.9999
Mean value (*MI*)	0.099	1.0000	30	0.033	1.0000
Distance along MST (*MST*)	5.264	0.9939	28	3.168	0.9979

Solvent polarity and feature reduction

All results reported so far were obtained from the complete data set of 158 solutes. It can be shown, however, that the same classification results if the data set is reduced to a much smaller number of solutes. This feature reduction process needs a ranking criterion that describes the retention characteristics of a solvent.

Different quantitative measures of the so-called solvent 'polarity' have been suggested[4,5,8,12,15] and a comparison of these methods has been made[23]. In all these approaches the data set of McReynolds[3] was used.

In this work known and new chromatographic solvent polarity characteristics were compared by means of regression analysis. In this approach the solvent polarity is defined by a linear combination of the retention indices of a number of representative solutes. These solutes are selected by the regression algorithm. In terms of pattern recognition, the solvent polarity is a continuous property, because there are no discrete categories of solvent polarities.

The measures of solvent polarity that will be compared are the following:

(1) The polarity constant (*MR*), defined by McReynolds[3]. Five of the ten solutes used originally for the calculation are included in the data set used in our work. The calculation of the McReynolds constant is carried out with these five solutes: (2-methyl-2-pentanol, 1,4-dioxane, benzene, 2-pentanone, butanol).

(2) The euclidian distance (*ED*) of the selected phase relative to the most non-polar stationary phase, *i.e.*, squalane.

(3) The mean of the retention indices (*MI*) of all key solutes.

(4) The length obtained by the traversal of the minimum spanning tree (*MST*).

The polarity values of all phases according to these definitions are shown in Table VI.

The calculation leading to the selection of representative solutes is carried out in a stepwise linear regression mode. Linear means that only linear terms are included in the model; stepwise means that only one additional solute is included in or excluded from each calculation step. The inclusion or exclusion of a solute is decided according to the significance level of this solute in the change of the variance of the error (*F*-test). The results are given in Table VII. As different numbers of solutes were used

TABLE VIII

SELECTION OF TEN SOLUTES FOR THE CLASSIFICATION OF SOLVENTS (FEATURE SELECTION)

Selection criterion	*Solutes selected*
McReynolds constant	3-Methyl-3-pentanol, methacrolein, 1,4-dioxane, 2-ethyl-1-butanol, benzene, 1,4-butylene glycol formal, 2-pentanone, butanol, methanol, toluene
Euclidian distance	Propyl acrylate, 1,3-propylene glycol formal, 1,3-butylene glycol formal, pentyl butyrate, 2-hexanol, tetrahydrofuran, 3-pentyl formate, chloroform, cyclohexyl acetate, propyl formate
Mean retention index	Methyl acrylate, cyclopentanone, 2-hexanol, isopropyl ethyl formal, 2,4-pentandione, 2,2-dimethyl-1-butanol, pentyl butyrate, 2-buten-1-ol, heptanal, 1,3,5-trioxepane
Distance along MST	2-Butanone, 1,3,5-trioxepane, butyl isobutyrate, diethyl propylal, 3-pentyl acetate, *tert.*-butyl acetate, 2,2-dimethylpropionaldehyde, 3,3-dimethyl-2-butanone, acrolein diethyl acetal, isobutyl isobutyrate

in the calculation of the different solvent polarity measures, all calculations were stopped when ten solutes had been selected. These ten most selected solutes are listed in Table VIII. It can be seen that the simplest criterion for the solvent polarity, the average retention index, gives the best results.

REFERENCES

1 L. Rohrschneider, *J. Chromatogr.*, 22 (1966) 23–28.
2 L. Rohrschneider, *Anal. Chem.*, 45 (1966) 1241–1247.
3 W. O. McReynolds, *J. Chromatogr. Sci.*, 8 (1970) 685–691.
4 J. J. Leary, J. B. Justice, S. Tsuge, S. R. Lowry and T. L. Isenhour, *J. Chromatogr. Sci.*, 11 (1973) 201–206.
5 S. R. Lowry, S. Tsuge, J. J. Leary and T. L. Isenhour, *J. Chromatogr. Sci.*, 12 (1974) 124–127.
6 J. K. Haken, M. S. Wainwright and N. Do Phuong, *J. Chromatogr.*, 117 (1976) 23–28.
7 D. L. Massart, P. Lenders and M. Lauwereys, *J. Chromatogr. Sci.*, 12 (1974) 617–625.
8 A. Eskes, F. Dupuis, A. Dijkstra, H. De Clercq and D. L. Massart, *Anal. Chem.*, 47 (1975) 2168–2174.
9 H. De Clercq, M. Despontin, L. Kaufman and D. L. Massart, *J. Chromatogr.*, 122 (1976) 535–551.
10 D. L. Massart and H. L. O. De Clercq, *Advan. Chromatogr.*, 16 (1978) 75–111.
11 J. O. De Beer and A. M. Heyndrickx, *J. Chromatogr.*, 235 (1982) 337–349.
12 S. Wold, *J. Chromatogr. Sci.*, 13 (1975) 525–532.
13 R. N. Carey, S. Wold and J. O. Westgard, *Anal. Chem.*, 47 (1975) 1824–1829.
14 D. H. McCloskey and S. J. Hawkes, *J. Chromatogr. Sci.*, 13 (1975) 1–5.
15 S. R. Lowry, G. L. Ritter, H. B. Woodruff and T. L. Isenhour, *J. Chromatogr. Sci.*, 14 (1976) 126–131.
16 P. H. Weiner and J. F. Parcher, *J. Chromatogr. Sci.*, 10 (1972) 612–615.
17 G. Dahlmann, H. J. K. Köser and H. H. Oelert, *J. Chromatogr. Sci.*, 17 (1979) 307–313.
18 J. K. Haken and D. Srisukh, *J. Chromatogr.*, 199 (1980) 199–208.
19 W. O. McReynolds, *Gas Chromatographic Retention Data*, Preston Technical Abstracts Company, Evanston, IL, 1966.
20 S. R. Heller and R. Potenzone, Jr., *Trends Anal. Chem.*, 2 (1983) 218–221.
21 D. L. Duewer, A. M. Harper, A. M. Koskinen, J. L. Fasching and B. R. Kowalski, *ARTHUR, Version 3-7-77* Infomatrix, Seattle, WA.
22 B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 95 (1973) 686–693.
23 S. R. Lowry, H. B. Woodruff and T. L. Isenhour, *J. Chromatogr. Sci.*, 14 (1976) 129–131.