Analysis of a system of description of odors by means of four different multivariate statistical methods

Maurice Chastrette, Jean-Yves de Saint Laumer and Philippe Sauvegrain[1]

Laboratoire de Chimie Organique Physique, Université Lyon 1, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France and [1]Firmenich SA, Case Postale 239, CH-1211, Genève 8, Switzerland

Abstract. In order to analyze the relationships among 32 descriptors of odors (notes), similarity coefficients were calculated using a data bank of 628 odoriferous products. The simple examination of the similarity matrix (32,32) has shown notes selectively and strongly associated (e.g. camphoraceous−piney and musky−powdery) and others less selectively associated (e.g. floral, green and herbaceous). This analysis was completed by four multivariate statistical methods. Non-linear mapping (NLM) proved to be more efficient than principal coordinates analysis for planar representation of olfactory notes, and has given results similar to those previously obtained using other data and other methods (similar disposition of notes around the central note 'floral'). Furthermore, the ascending hierarchical taxonomy and the minimal spanning tree were coherent with the NLM representation. These three methods complete each other and constitute a convenient system to analyze odor descriptions.

Introduction

Perfumers use a relatively small number of words or notes to describe odors (Jaubert *et al.*, 1987; Randebrock, 1985; Elmouaffek, 1988). In a previous paper Chastrette *et al.* (1988) analyzed 24 notes frequently used by Arctander (1969) to describe odors. Using principal coordinates analysis, based on similarities among notes, they have shown that the corresponding olfactory space was not strongly structured. They distinguished three kinds of notes: those which are isolated (no strong association), those which are selectively associated (one or two strong associations) and those which are associated with a large number of notes (several associations). Isolated notes are favorable to the structure−activity analysis, unlike those which are not selectively associated and which correspond to a large range of chemical structures.

In this paper, another system of odor description, designed by Firmenich SA, is analyzed by means of four different statistical methods: principal coordinates analysis (PCA), (Massart and Kaufman, 1983), non-linear mapping (NLM) (Sammon, 1969), ascending hierarchical taxonomy (AHT) (Dagnelie, 1984) and minimal spanning tree (MST) (Lee *et al.*, 1977). The two objectives of these analyses are: (i) to confirm or modify the results concerning Arctander's odor description system (1969), or another description system, and (ii) to compare the four statistical methods and find those which are the most consistent, in agreement with the empirical rules of perfumery.

Materials and methods

Data

A data bank of 628 pure odorous substances, commonly encountered in perfumery and compiled by Firmenich SA, has been analyzed. Each product is described by a team of seven perfumers by two, three or four words called notes and chosen among 32

possible descriptors. The descriptors, which may be repeated, are cited in order of decreasing intensity. Only the first three descriptors have been used in the present work.

Construction of the similarity matrix

The 628 pure substances of the data bank are considered as individual entities defined by the 32 variables which are associated with notes and which take the values 1 or 0.

Let $X(32 \times 628)$ be the data matrix. The value of $x_{i,j}$ corresponding to the jth compound and to the ith note, is either 1 if the note 'i' is present in the description of the jth compound or 0 otherwise. Then, the matrix of occurrence/co-occurrence $C(32 \times 32)$ is calculated as the product of the matrix X by its transpose matrix ^tX.

From matrix C, the similarity matrix $S(32 \times 32)$ is calculated using several similarity coefficients proposed in the literature (Iglesias, 1975; Sokal and Sneath, 1963). In a recent study, Chastrette *et al.* (1988) have selected the similarity coefficient proposed by Ochiai (1957) as it does not take into account a situation where the substances present neither the i note nor the j note. This similarity coefficient, $s_{i,j}$, which is an element of the similarity matrix S, is then defined by:

$$s_{i,j} = c_{i,j}/(c_{i,i} \times c_{j,j})^{1/2}, \quad 0 \leq s_{i,j} \leq 1$$
$$c_{i,j} = \text{element of the matrix of occurrence/co-occurrence } C$$

Test of significance of associations

The analysis of similarity between two notes i and j must be completed by a test of significance. Such a test in a 2×2 contingency table is easily performed (Lamont, 1949), by means of the χ^2 test, with a single degree of freedom. For each couple of notes, we determine a similarity coefficient $s_{i,j}$ and a χ^2 value which measures the significance of this similarity. The level of 90% significance corresponds to a χ^2 value equal to 2.7 (Wine, 1964).

Different methods of analysis

A: principal coordinates analysis (PCA). The similarity matrix S is transformed into a matrix T suitable for PCA (Gower, 1966), according to the equation:

$$t_{i,j} = s_{i,j} - s_i - s_j + s$$

where $t_{i,j}$ = element of the transformed matrix T; $s_{i,j}$ = element of the similarity matrix S; s_i = mean of row i; s_j = mean of column j; s = mean of all the elements of the matrix S.

During a second stage, PCA is performed on the transformed matrix T and several factorial planes are used to represent the relationships among the 32 descriptors used in the data bank.

B: non-linear mapping (NLM). Multidimensional scaling has been used by Schiffman (1974) to classify odorous chemical substances and to represent them on a map. Here, we have used a slightly different method, the non-linear mapping (NLM), to obtain a planar representation of the descriptors. In both methods, projections are made in order to preserve the distances between descriptors, considered as points in an L-dimensional space, as well as possible.

Sammon (1969) has described an algorithm based on a point mapping of N L-dimensional vectors from the L-space to a lower-dimensional space so that the inherent data 'structure' is approximately preserved.

Let X_i ($i = 1, \ldots, N$) be N points defined by L coordinates in an L-space and let the distance between the points X_i and X_j in the L-space be defined by $d_{i,j}^* = \text{dist}(X_i, X_j)$.

Let Y_i ($i = 1, \ldots, N$) be N points corresponding to the X_i points and defined by D coordinates in a D-space (D = 2 or D = 3) and the distance between the points Y_i and Y_j in the D-space be defined by $d_{i,j} = \text{dist}(Y_i, Y_j)$.

An initial configuration is randomly chosen for the Y_i points. Next, all the D-space distances $d_{i,j}$ are computed and used to define an error E. E indicates how well the present configuration of the N points in the D-space fits the N points in the L-space.

$$E = \frac{1}{c} \sum_{i<j}^{N} [d_{i,j}^* - d_{i,j}]^2 / d_{i,j}^*$$

where $c = \sum_{i<j} [d_{i,j}^*]$

The next step in the NLM algorithm consists of adjusting the $Y_{p,q}$ coordinates of the points Y_i to decrease the error E. A steepest descent procedure is used to search for a minimum error.

As this method is based on distances, the similarity matrix **S** has been transformed into a distance matrix **D** in the following way:

$$d_{i,j}^* = ((1/s_{i,j}) - 1)/100 \text{ and } d_{i,j}^* = 1 \text{ if } s_{i,j} < 0.01$$

This relation has been chosen to ensure that the ratio between the larger and the smaller distances is significant ($\simeq 100$) and that the larger distances are limited to 1 in order to avoid too strong an influence of those distances on the NLM construction.

C: ascending hierarchical taxonomy (AHT). In hierarchical clustering methods, several clusters of a set of objects are obtained and are ranked in such a way that smaller clusters are included in larger clusters.

In AHT (Benzecri, 1980) one starts with *m* objects to be classified, and at each step the two objects (or clusters) which are the most similar are merged into a single cluster. After $m - 1$ such steps, all objects belong to one large cluster. There are many AHT methods, differing in the criteria used to decide which individual elements or clusters should be merged and the way in which the similarity between a newly obtained cluster and other clusters or objects is defined.

Let A,B, C be three objects or clusters and G a cluster constituted by A and B. Let d(X,Y) be the distance between X and Y. Let nA and nB be respectively the number of objects of the clusters A and B. The three aggregation methods commonly used are the following:

Single linkage: $d(G,C) = \min(d(A,C), d(B,C))$

Complete linkage: $d(G,C) = \max(d(A,C), d(B,C))$

Average linkage: $d(G,C) = \dfrac{[nA \times d(A,C) + nB \times d(B,C)]}{nA + nB}$

The distance between two i and j notes is usually taken as $1 - s_{i,j}$ (Dagnelie, 1984). The structure of the classification can be represented graphically by linking the elements together in a dendrogram.

D: minimal spanning tree (MST). This method gives a graph in which no attempt is made to preserve all the distances. Only the distances to the nearest neighbors are preserved accurately.

Given a set of points, a spanning tree is a connected graph that satisfies the following conditions: (i) every point is on the graph, (ii) there is no loop on the graph, and (iii) the MST is a tree, the total length of which is minimum.

If there are no two links of the MST with equal length, there is only one MST solution. This solution only contains all the information about the distances of the nearest neighbor of every point.

Lee *et al.* (1977) have proposed building this spanning tree using a triangulation method in which the distance of each point on the graph to two previously mapped points is exactly preserved.

Let Pi and Pj be points already mapped. Let Pk be directly linked to Pj on the MST. Pk is mapped on the 2-space in such a way that its distances to Pi and Pj are exactly preserved. Pi is either the point linked to Pj and preceding Pj in the MST (second nearest neighbor approach) or a central reference point (reference point method).

It is impossible to use the same distance calculation as in NLM because of the triangular inequality (it is not always possible to draw three points while keeping their relative distances). In order to obtain a more readable representation, differences between the distances have been increased by transformation of the usual distance calculation: $d_{i,j} = 0.55 - S_{i,j}$ instead of $d_{i,j} = 1 - S_{i,j}$. These transformations do not modify either the distance ranks or the links between notes. As in this case the reference point method has not proved to be suitable, owing to the large number of notes (32), we have used the second nearest neighbor approach for the MST construction.

Results

Analysis of similarities

To analyze the similarities among descriptors we have to consider three factors: the similarity coefficient, the χ^2 value and the total number of occurrences for each possible association. The $S_{i,j}$ coefficients, elements of the matrix S (Table I), are often very low. If the limit of similarity is set to $s \geq 0.08$ [as in (4)], only 107 associations (21.5%) have to be considered. However, only 26 are significant at a 90% level, i.e. have $\chi^2 \geq 2.7$.

In short, an association corresponding to high values of s and χ^2 is both strong and significant. Examples of this kind are the camphoraceous−piney ($s = 0.48$; $\chi^2 = 72.9$) and ethereal−fruity ($s = 0.43$; $\chi^2 = 65.3$) pairs.

On the contrary, some pairs correspond to lower values of s and χ^2 and can be considered as not associated. For example, while each descriptor could be involved in 31 associations, musky has 27 association coefficients equal to 0, and amber has 22. Their rare associations are often very significant. Actually, musk is associated

Table I. Similarity matrix S calculated using Ochiai's coefficient

	ANISIC	ETHEREAL	FLORAL	SPICY	FRUITY	CAMPHORACEOUS	PINEY	POWDERY	FATTY	CITRUS	HERBACEOUS	MINTY	WOODY	CULINARY	SULFURY	BALSAMIC	ROSE	HUMUS	GREEN	ALMOND	VIOLET	PHENOLIC	ACIDIC	SALICYLATE	ORANGE-FLOWER	LACTONIC	HAY	HONEY	HUSKY	ANIMAL	AMBER	PYROGENOUS
ANISIC	1.00																															
ETHEREAL	0.03	1.00																														
FLORAL	0.23	0.05	1.00																													
SPICY	0.00	0.00	0.18	1.00																												
FRUITY	0.07	0.43	0.21	0.02	1.00																											
CAMPHORACEOUS	0.00	0.03	0.05	0.03	0.05	1.00																										
PINEY	0.00	0.02	0.00	0.00	0.03	0.48	1.00																									
POWDERY	0.17	0.00	0.25	0.18	0.06	0.00	0.06	1.00																								
FATTY	0.00	0.10	0.18	0.07	0.09	0.04	0.02	0.05	1.00																							
CITRUS	0.02	0.03	0.19	0.07	0.06	0.04	0.14	0.01	0.21	1.00																						
HERBACEOUS	0.13	0.00	0.20	0.07	0.15	0.22	0.15	0.06	0.18	0.28	1.00																					
MINTY	0.05	0.03	0.05	0.00	0.02	0.18	0.00	0.03	0.00	0.00	0.23	1.00																				
WOODY	0.00	0.02	0.07	0.12	0.00	0.09	0.15	0.23	0.00	0.04	0.01	0.00	1.00																			
CULINARY	0.03	0.00	0.09	0.09	0.10	0.00	0.00	0.04	0.07	0.00	0.09	0.00	0.05	1.00																		
SULFURY	0.00	0.00	0.02	0.00	0.12	0.00	0.05	0.00	0.00	0.03	0.05	0.00	0.00	0.23	1.00																	
BALSAMIC	0.00	0.02	0.21	0.26	0.09	0.02	0.03	0.23	0.00	0.02	0.01	0.00	0.07	0.00	0.00	1.00																
ROSE	0.00	0.00	0.10	0.00	0.18	0.00	0.09	0.07	0.05	0.17	0.18	0.00	0.00	0.00	0.00	0.08	1.00															
HUMUS	0.03	0.05	0.18	0.05	0.00	0.15	0.18	0.04	0.03	0.03	0.18	0.00	0.15	0.02	0.08	0.02	0.08	1.00														
GREEN	0.05	0.14	0.26	0.03	0.24	0.05	0.02	0.04	0.19	0.12	0.27	0.02	0.01	0.18	0.07	0.08	0.11	0.22	1.00													
ALMOND	0.14	0.00	0.09	0.00	0.04	0.00	0.00	0.05	0.05	0.00	0.04	0.00	0.00	0.12	0.00	0.05	0.03	0.00	0.08	1.00												
VIOLET	0.14	0.00	0.09	0.00	0.05	0.05	0.10	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.08	0.00		1.00											
PHENOLIC	0.04	0.02	0.07	0.07	0.01	0.03	0.03	0.07	0.00	0.02	0.10	0.00	0.03	0.00	0.00	0.05	0.00	0.10	0.02	0.04	0.00	1.00										
ACIDIC	0.00	0.00	0.02	0.00	0.15	0.00	0.00	0.02	0.10	0.05	0.05	0.00	0.03	0.00	0.05	0.03	0.05	0.00	0.12	0.00	0.00	0.00	1.00									
SALICYLATE	0.00	0.04	0.18	0.00	0.11	0.00	0.04	0.04	0.00	0.05	0.00	0.02	0.00	0.00	0.07	0.05	0.02	0.08	0.04	0.04	0.21	0.00		1.00								
ORANGE-FLOWER	0.00	0.00	0.05	0.00	0.02	0.00	0.00	0.10	0.03	0.06	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.11	0.00	0.00	0.04	0.00	0.00	0.17	1.00							
LACTONIC	0.00	0.03	0.05	0.00	0.11	0.00	0.00	0.07	0.05	0.00	0.08	0.00	0.03	0.18	0.00	0.08	0.00	0.00	0.00	0.05	0.00	0.00	0.05	0.04	0.00	1.00						
HAY	0.05	0.00	0.02	0.00	0.15	0.00	0.08	0.00	0.03	0.05	0.20	0.00	0.09	0.00	0.04	0.00	0.07	0.05	0.08	0.08	0.14	0.00	0.00	0.00	0.17		1.00					
HONEY	0.04	0.03	0.18	0.00	0.09	0.04	0.00	0.02	0.03	0.02	0.00	0.00	0.00	0.08	0.17	0.08	0.00	0.00	0.00	0.05	0.07	0.00	0.00	0.00				1.00				
HUSKY	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						1.00			
ANIMAL	0.33	0.00	0.15	0.03	0.03	0.00	0.00	0.14	0.09	0.04	0.02	0.00	0.00	0.03	0.05	0.00	0.02	0.05	0.03	0.00	0.00	0.28	0.00	0.03	0.00	0.00	0.05	0.07	0.18	1.00		
AMBER	0.00	0.00	0.04	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.32	0.00	0.00	0.04	0.00	0.04	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	1.00	
PYROGENOUS	0.00	0.00	0.02	0.05	0.07	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.19	0.08	0.00	0.00	0.00	0.00	0.28	0.00	0.10	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

significantly only with the note powdery (s = 0.33; χ^2 = 95.1), and amber with animal (s = 0.19; χ^2 = 3.72) and woody (s = 0.32; χ^2 = 55).

Between these extreme cases, descriptors such as floral, fruity, herbaceous and green present numerous associations when only si,j is taken into account. For example, floral appears as associated with 17 other descriptors (15 for fruity, 13 for herbaceous and 14 for green). The four descriptors are involved in 59 of the 107 pairs with s ≥ 0.08. However, these apparent associations are produced by the high number of occurrences and consequently of co-occurrences of those notes. In fact, floral has significant associations with anisic only (s = 0.25; χ^2 = 7.0).

These observations can be compared with the conclusions of a previous study (Chastrette *et al.*, 1986) in which 24 notes were analyzed by means of PCA and AHT, from the description of 2467 chemical substances by Arctander (1969). A similarity matrix was used in a PCA and in an AHT. The main results were as follows. The camphoraceous − minty pair has the strongest association (in our study s = 0.18 and χ^2 = 7.82). The most frequently associated notes were floral, fruity and green. The less frequently associated notes were musky, amber and anisic. In this previous study, the notes powdery and piney were not analyzed because they were not used in the odor description by Arctander.

Principal coordinates analysis

The principal coordinates analysis (PCA) of the matrix T (the latter is the result of the transformation of the similarity matrix S) gives a representation of the olfactory space defined by the 32 notes of the Firmenich data bank. The results of the PCA are given in Table II.

The eigenvalues decrease slowly, and the first two axes account for only 11.64% of the total variance. Five axes are needed to represent 25% of the information.

If all the factors were equivalent, each of them would represent ∼ 3% of total variance. Here the total variance contributed by the first axis is only 6%. These results reflect the loose structure of the olfactory space. A similar study of Arctander's data has led to the same conclusion (Chastrette *et al.*, 1986). Figure 1 shows the projection of the olfactory notes on the first two factorial axes. This graph brings out associations between camphoraceous − piney, animal − musky, green − fruity, woody − amber. Such associations are interesting, but the very small percentage of total variance contributed by each axis does not allow us to use this representation only. The other projections on factorial axes 1 − 3 or 2 − 3 are not presented here because none of them gives interpretable structures. Each of these representations accounts for a limited percentage of the total variance. In all these projections, some associations are preserved while

Table II. Percentage of total variance contributed by each eigenvalue

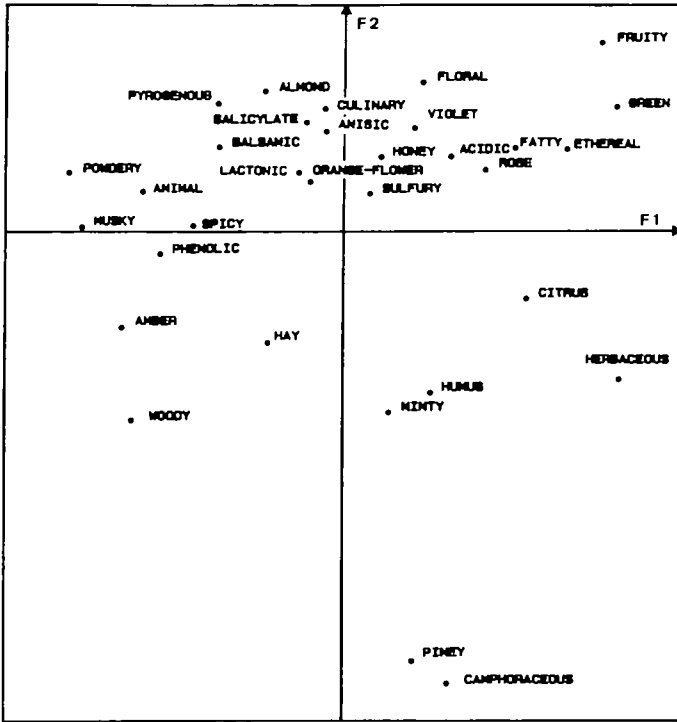| Factor	Eigenvalue	Percentage of total variance	Cumulated percentage
1	1.75	5.95	5.95
2	1.67	5.68	11.64
3	1.51	5.12	16.76
4	1.36	4.62	21.38
5	1.32	4.48	25.87

Fig. 1. Projection of the 32 olfactory notes on the first two factorial axes (F1 and F2).
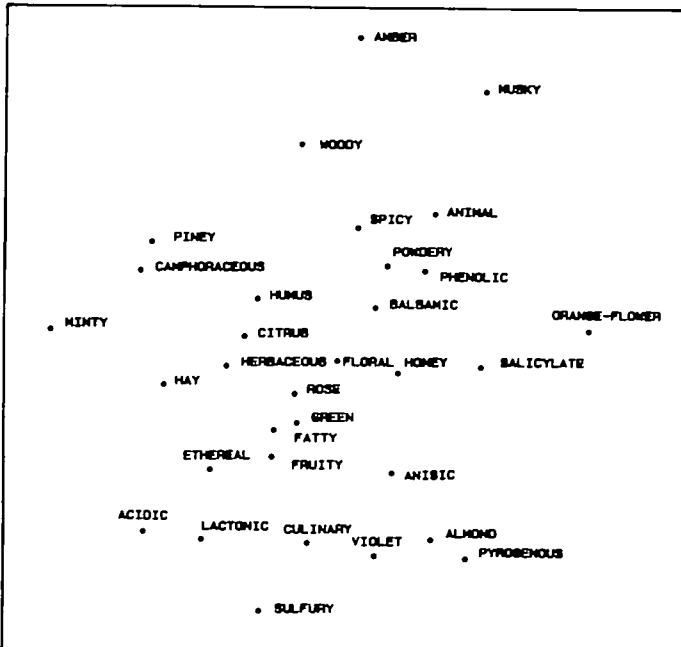

Fig. 2. Non-linear mapping of the 32 olfactory notes.

others are not shown. For these reasons it is difficult to distinguish the real associations from the superpositions due to projection and PCA appears to be poorly adapted to our problem.

Non-linear mapping

Non-linear mapping (NLM) is also a global analysis and gives graphic solutions after a quick convergence of the error coefficient E (often less than 300 iterations).

Two-hundred random initial configurations were used in the NLM algorithm. All the solutions which were obtained after 300 iterations are very similar and are representative of the local minima of the E coefficient. The solution presented in Figure 2 was obtained many times (allowing for rotation and symmetry) and corresponds to the best solution according to the criterion defined by NLM.

Comparison between PCA and NLM

PCA shows some of the strongest associations, but the positions of some notes are incorrect. So, one of the less associated notes, anisic, appears in the middle of Figure 1, clustered with many notes. On the contrary, floral, the most frequently associated note, is far from the center of the figure.

The low association of anisic is represented by its isolated position on the NLM (Figure 2) and most of the perfumers' remarks are respected by the NLM: the minty − camphoraceous − piney − woody series is easily seen on the NLM but this is not the case on the PCA graph. The clustering of floral, honey and rose notes constitutes a central nucleus. On the contrary, in the PCA representation, acidic is inside a triangle constituted by these notes. This is contrary to the perfumers' experience. According
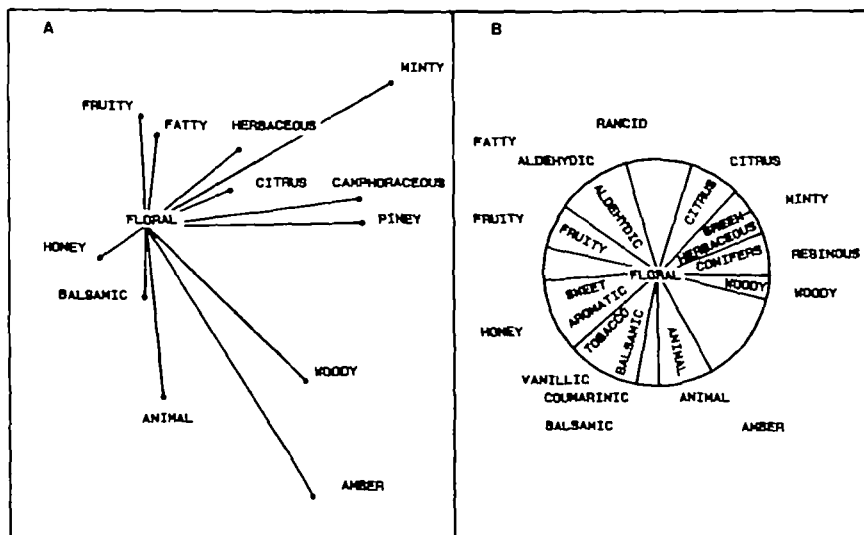

Fig. 3. Representation of the olfactory space centered on the note floral. (A) From our results. (B) Inside, from Haarman and Reimer. (C) Ouside, from Jaubert.

to perfumers, acidic can be considered as the next element of the series fruity — ethereal, as represented on the NLM.

The only disagreement between perfumers and the NLM graph (Figure 2) is the
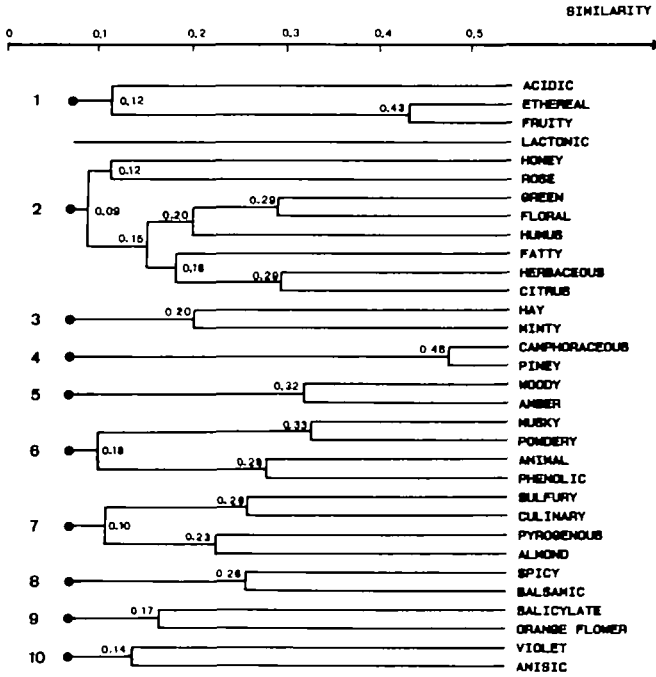

Fig. 4. Dendrogram of the 32 olfactory notes obtained using mean aggregation techniques.
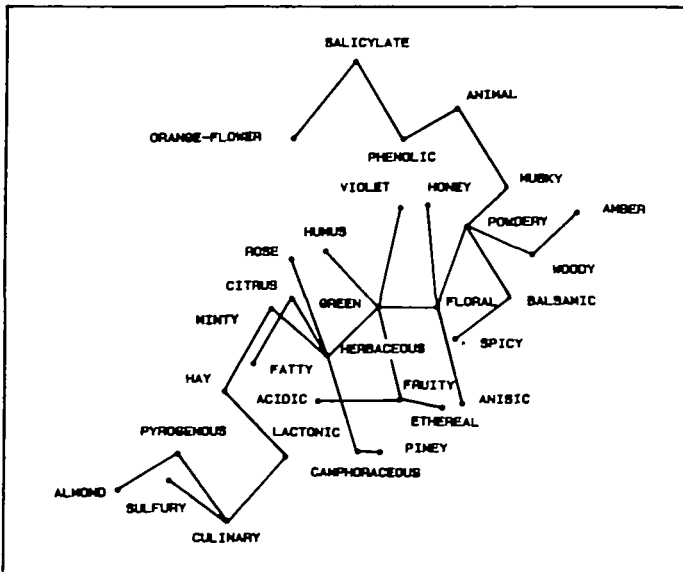

Fig. 5. Minimal spanning tree of the 32 olfactory notes obtained using the second nearest neighbor method.

position of some notes of flowers such as orange-flower and violet which are a long way from the floral note. This could be explained by the fact that the floral note is not used to describe odors such as orange-flower and violet.

The NLM is more compatible with the perfumers' observations than the PCA.

The differences between both methods could result from the differences in computation of distances. To test this hypothesis we performed a PCA using the same distance calculations as in NLM. We have found that the PCA does not give better results than previously.

We have observed that floral has a central position on the NLM. Haarman and Reimer (1979) using empirical results, and Jaubert *et al.* (1987) using PCA, have obtained circular graphs (perfumers' rosace) representing the olfactory space centered on the same floral note. In order to compare these graphs with our results we drew a graph based on NLM (Figure 3). We have noticed that the order of the notes in the three representations is the same, except for the minty−citrus inversion. These results are similar to a perfumers' rosace drawn a few years ago by Firmenich SA perfumers.

Ascending hierarchical taxonomy

Analysis of similarity matrix S by ascending hierarchical taxonomy (AHT) gives two dendrograms using two aggregation techniques (Elmouaffek, 1988). The graph (Figure 4) obtained by the mean aggregation techniques allows us to divide notes into ten groups comprising two to eight notes. Only the lactonic note is not included in a group. These groups are the following:

1. Acidic, ethereal, fruity
2. Herbaceous, citrus, fatty, humus, green, floral, rose, honey
3. Minty, hay
4. Camphoraceous, piney
5. Amber, woody
6. Phenolic, animal, musky, powdery
7. Culinary, sulfury, pyrogenous, almond
8. Balsamic, spicy
9. Orange-flower, salicylate
10. Violet, anisic

Figure 4 shows that these groups do not present a strong hierarchy and confirms the results obtained in a previous AHT (Chastrette *et al.*, 1988; Elmouaffek, 1988), as well as PCA and NLM.

Minimal spanning tree

The minimal spanning tree (MST) has been drawn using the second nearest neighbor approach and starting from green as the central note because this note is linked to the three most frequently associated notes (floral, fruity and herbaceous). This choice has given the best legibility on the graph (Figure 5).

On this figure, the 14 distances between green and its five neighbors (violet, humus, fruity, floral, herbaceous) and the neighbors of these notes (honey, powdery, anisic, ethereal, acid, citrus, rose, minty, camphoraceous) are shown without deformation (i.e.

14 notes from the 32 analyzed). However, the MST gives no information about the distances between green and the other 17 notes.

A combination of AHT, MST and NLM

The results obtained by the three methods are different but complete each other. The clustering of notes obtained by AHT, completes the NLM results as shown in Figure 6. A group of eight strongly associated notes is located in the middle of the graph. Around them are the more selectively associated notes such as camphoraceous − piney − amber or phenolic − animal − musky. The observed agreement between two fundamentally different methods increases the credibility of the results obtained by both methods. On the whole, our groups correspond to the perfumers' point of view. However, to be in exact agreement with them, the sulfury − culinary − almond − pyrogenous group should be separated in two, as suggested by the NLM. In the same way, perfumers would put the lactonic note in the fruity − ethereal group. On the contrary, the violet − anisic association is not satisfactory but can be explained by the low occurrence of the violet note (18) which is associated with anisic because of a co-occurrence of only 3. Here we are limited by the data bank size.

MST does not allow us to distribute notes into groups but, for each note, it describes its close neighborhood correctly. Figure 7 shows the MST drawn on the NLM graph. This superposition leads to a very consistent system which shows the privileged relationship between notes and groups of notes clustered using the AHT techniques. The fruity − ethereal − acidic group is linked to the tree by the green note. The three
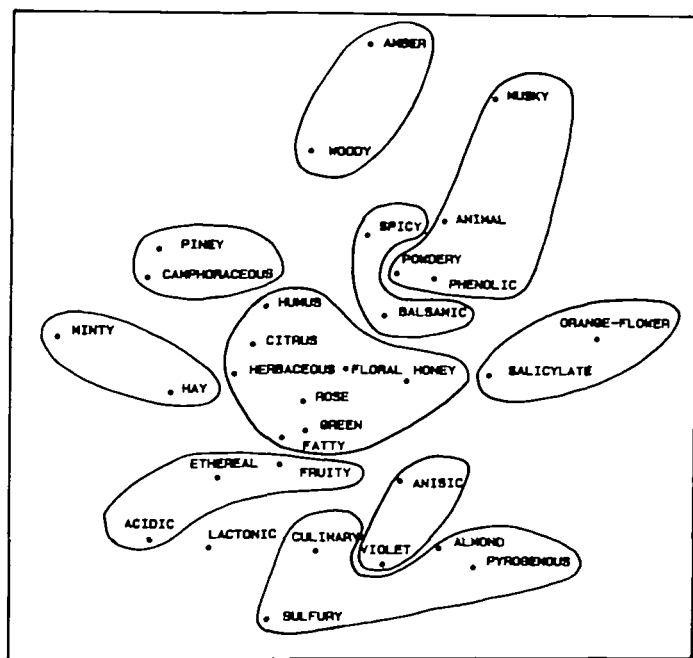

Fig. 6. The 10 groups of olfactory notes obtained using ascending hierarchical taxonomy drawn on the non-linear map.
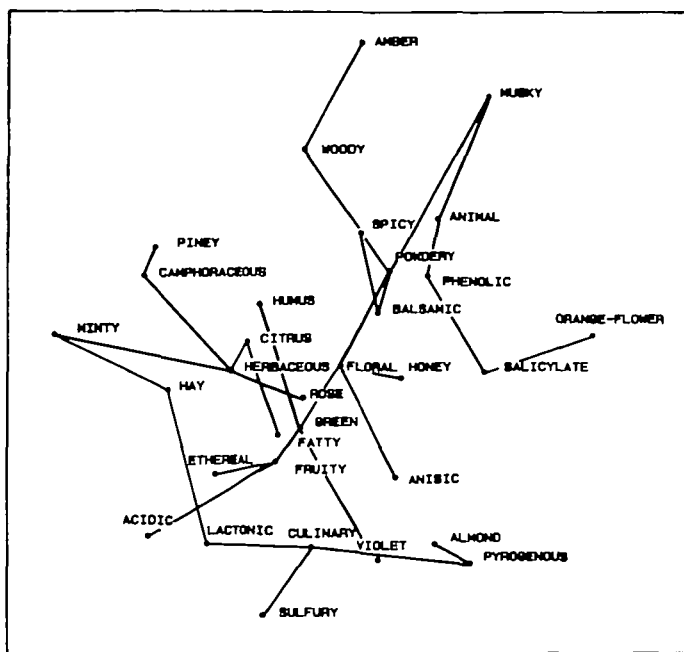
Fig. 7. Linkages of the minimal spanning tree drawn on the non-linear map.

amber – spicy – balsamic and musky – animal – phenolic – powdery groups are all linked to the same note floral. Minty – hay and camphoraceous – piney are linked to the tree by the herbaceous note. On the contrary, the anisic – violet group is not respected by the MST. This is consistent with the perfumers' remarks.

Conclusion

Among the four reported methods, PCA is obviously the least suitable for analyzing the relationships among descriptors of odors, due to the low percentage of total variance accounted for by the first axes. This is a consequence of the weak structure of the olfactory space in the Firmenich bank, previously observed in Arctander's description system. On the contrary, the three other methods are coherent and complete each other as shown in Figures 6 and 7 (except for local aberrations for the lactonic, violet and orange-flower notes).

NLM leads to results similar to those previously obtained by Jaubert *et al.* (1987), Haarman and Reimer (1979) and Chastrette *et al.* (1988) using other data and other methods.

The association of these three methods constitutes a convenient system for analysis of odor descriptions. A study of a much larger data bank is in progress to improve our results for the notes which have too small an occurrence in this bank.

Acknowledgements

References

Arctander,S. (1969) *Perfume and flavor chemicals.* Montclair, NJ.

Benzecri,J.P. (1980) *L'analyse des données. I. la taxinomie.* Dunod, Paris.

Chastrette,M., Elmouaffek,A. and Zakarya,D. (1986) Etude statistique multidimensionnelle des similarités entre 24 notes utilisées en parfumerie. *C. R. Acad. Sci. Paris,* **303**, Série II, 1209−1214.

Chastrette,M., Elmouaffek,A. and Sauvegrain,P. (1988) A statistical study of notes used in perfumery. *Chem. Senses,* **13**, 295−305.

Dagnelie,P. (1984) *Analyse statistique à plusieurs variables.* Presses Agronomiques de Gembloux, Belgique.

Elmouaffek,A. (1988) Etude statistique multidimensionnelle des proximités entre notes odorantes utilisées en parfumerie. Lyon, Thèse de doctorat, no. 56-88.L.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika,* **53**, 325−338.

Haarman and Reimer (1979) Le discodor. *H et R Contact,* **23**, 19.

Iglesias,A. (1975) Contribution à l'analyse des tableaux de distances et de similarité. Application à la biologie. Lyon, Thèse de 3ᵉ cycle, no. 444.

Jaubert,J.N., Gordon,G. and Dore,J.C. (1987) Une organisation du champ des odeurs. *Parf. Cosm. Aröm.,* **78**, 71−81.

Lamont,C.C. (1949) The measurement of interspecific association. *Ecology,* **30**, 411−424.

Lee,R.C.T., Slagle,J.R. and Blum,H. (1977) A triangulation method for sequential mapping of points from N-space to two-space. *IEEE Trans. Comput.,* A-25, 288−292.

Massart,D.L. and Kaufman,L. (1983) *The interpretation of analytical chemical data by the use of cluster analysis.* Wiley, New York.

Ochiai,A. (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bull. Jap. Sci. Fish,* **22**, 526−530.

Randebrock,R.E. (1985) Geruch und Konstitution: Eine Beweisführung für die Molekulartheorie des Geruches. *Parfüm. Kosmet.,* **66**, 435−443.

Sammon,J.W. (1969) A non-linear mapping for data structure analysis. *IEEE. Trans. Comput.,* c-18, 401−409.

Schiffman,S.S. (1974) Physicochemical correlates of olfactory quality. *Science,* **185**, 112−117.

Sokal,R.R. and Sneath,P. (1963) *Principles of numerical taxonomy.* Freeman, San Francisco.

Wine,R.L. (1964) *Statistics for scientists and engineers.* Prentice-Hall, Englewood Cliffs, NJ.