Identification of Latent Variables in a Semantic Odor Profile Database Using Principal Component Analysis

Manuel Zarzo and David T. Stanton

Corporate Research, Modeling, and Simulations Department, Procter & Gamble Co., Miami Valley Innovation Center, 11810 East Miami River Road, Cincinnati, OH 45252, USA

Correspondence to be sent to: Manuel Zarzo, Corporate Research, Modeling and Simulations Department, Procter & Gamble Co., Miami Valley Innovation Center, 11810 East Miami River Road, Cincinnati, OH 45252, USA. e-mail: zarzo.mz@pg.com

Abstract

Many classifications of odors have been proposed, but none of them have yet gained wide acceptance. Odor sensation is usually described by means of odor character descriptors. If these semantic profiles are obtained for a large diversity of compounds, the resulting database can be considered representative of odor perception space. Few of these comprehensive databases are publicly available, being a valuable source of information for fragrance research. Their statistical analysis has revealed that the underlying structure of odor space is high dimensional and not governed by a few primary odors. In a new effort to study the underlying sensory dimensions of the multivariate olfactory perception space, we have applied principal component analysis to a database of 881 perfume materials with semantic profiles comprising 82 odor descriptors. The relationships identified between the descriptors are consistent with those reported in similar studies and have allowed their classification into 17 odor classes.

Key words: cluster, dimension, odor classification, odor descriptor, semantic profile

Introduction

Since the discovery of the large family of genes encoding putative olfactory receptors (ORs) (Buck and Axel 1991), many efforts have been successfully conducted to unravel the molecular and physiological basis of olfaction, but many details are still poorly understood. The sense of vision is governed by 3 types of receptors, and we can find 2 objects with exactly the same color. But olfaction involves a few hundred ORs, and it is believed that no 2 odorants have exactly the same odor (Turin and Yoshii 2003). Thus, humans can recognize or discriminate thousands of different odors. The reason seems to be that one odorant can activate different ORs and one OR recognizes multiple odorants, resulting in a combinatorial receptor-coding scheme to encode odor identities (Malnic et al. 1999).

Odor description

Smell is a sensation that is difficult to describe, measure, and predict, and hence, perfume research is still rather empirical. Attempting to provide certain standards in fragrance technology, perfumers have tried for many decades to develop an accurate description of odors. To characterize odor profiles, one option is to rate the smell similarity by direct comparison with a series of reference odorants (Schutz 1964;

Yoshida 1975). This is an objective approach but becomes time consuming and impractical with a high number of references. On the contrary, semantic methods allow for the rapid generation of data and consequently are the most commonly used procedures. They consist of assigning the words that come to mind when smelling a substance. Our odor memory "compares" the perceived sensation with those of other substances previously smelled. If there is a good match, one word can be enough, but usually several are necessary to describe how the smell resembles other common odors. These words are called odor character descriptors or notes. The most useful ones in fragrance chemistry are those generally understood and are usually associated with the source of that smell, making it easy for any observer to use after some training. An open or unrestricted description of a smell usually produces subjective characterizations such as "dry," "fresh," "powerful," "rich," "feminine," "natural," "tender," or "warm," etc. Their use should be avoided since they reflect one person's opinion and are open to discussion. In order to generate certain consensus, panelists are usually requested to assign for a given odorant those objective descriptors that best apply from a fixed list (Harper et al. 1968; Moskowitz and Barbe 1977). Because the use of verbal

odor descriptors requires observers to assign the same words in the same way, training and some experience are required. Although semantic methods have been considered significantly "noisy" because of interindividual differences in the interpretation of descriptors, the use of a panel provides an average odor profile that tends to stabilize if a large number of panelists are used (Dravnieks 1982). Moreover, although reference-odorant methods seem a priori more accurate, an experiment conducted with 49 panelists revealed that semantic methods were almost as reproducible as in direct comparisons (Dravnieks et al. 1978).

Odor classification

Although for an inexperienced observer the large list of descriptors for odor profiling seems to reflect a high dimensionality of odor space, some of the descriptors are related, and the basic smell attributes are easily identified after some training. Understanding the different relationships, associations, or similarities between these notes is the basis to define more accurately the olfactory universe of perfumers. Additionally, it may provide some insight into the basis of olfaction.

The first scientific approach for odor classification was proposed by Linnaeus based on his deep experience in botany (Linnaeus 1756). This 7-category system was revised later, and 2 new classes were added (Zwaardemaker 1925). For many decades, researchers have proposed a relatively small number of odor classes or dimensions in odor space, ranging from 4 to 9 (Henning 1916; Lovell 1923; Crocker and Henderson 1927; Klein 1947; Amoore 1962; Schutz 1964). Conversely, other authors consider that there are likely to be >20 descriptive terms that are essential to cover the complete range of odor stimuli (Harper et al. 1968). A classification employing 45 groups has also been proposed (Cerbelaud 1951). Many other efforts have been conducted to develop a consensus for odor classification (Woskow 1968; Kastner 1973; Yoshida 1975; Schiffman 1981; Jaubert et al. 1986; Lawless 1988, 1993; Higuchi et al. 2004), but none of them have yet gained wide acceptance.

Odor profile databases

Over time, fragrance chemical companies have developed databases of thousands of odorants with their corresponding odor profiles. Unfortunately, most of these data are not available to the scientific community, seriously limiting efforts to develop an accurate characterization of odor space. Such information would be beneficial for providing a means for the development of new odorants. But despite many efforts in obtaining structure–odor relationships (SORs) that may guide a rational approach for odorant discovery (Rossiter 1996), this goal is still mainly achieved by trial and error.

The most comprehensive published databases of odor profiles are the Arctander's handbook (Arctander 1969) and the Fenaroli's handbook (Burdock 2004). From both sources, a set of 1396 pure substances was compiled and analyzed, leading to a descriptive model of olfactory perception space (Jaubert et al. 1986, 1987). Arctander's handbook contains the odor description of 3102 perfume and flavor chemicals. Considering that the perfumer's palette now consists of approximately 4000 raw materials, this is a valuable reference for perfumers and flavorists. But odors have been basically characterized by only one person (S Arctander), resulting in an arguable degree of personal subjectivity. Moreover, many chemical structure drawings are not accurate. In total, about 270 odor descriptors are used, many of them subjective. In a first attempt to identify associations among these descriptors, a reported study (Chastrette et al. 1986) selected 24 notes and analyzed them using principal component analysis (PCA) and ascending hierarchical taxonomy (AHT). In a further effort to analyze this database, 74 notes were selected for 2467 pure substances (Chastrette et al. 1988). After calculating the similarity for every pair of descriptors, 2 hierarchical agglomerative classification methods were applied to identify statistically significant associations. As a result, 60 notes were regrouped in 27 clusters, each containing 2–4 notes, and 14 remained as isolated notes. In another study of the Arctander's handbook, 126 odor descriptors were selected for 1573 compounds, and a cluster analysis resulted in 19 clusters (Abe et al. 1990).

In order to assess the reproducibility of these results, a similar analysis was performed in a later study (Chastrette et al. 1991) of another database of 628 pure compounds compiled by SA Firmenich, La Plaine, Switzerland. Each product was described by a team of 7 perfumers, who assigned 2–4 notes chosen among 32 possible descriptors, and the 3 most frequent ones were considered as the odor profile. A similarity matrix was calculated as in the previous case (Chastrette et al. 1988) and was analyzed using 4 multivariate methods: nonlinear mapping, AHT, minimal spanning trees, and PCA. Several clusters of descriptors were obtained, consistent with the perfumers' point of view. Although this odor profile database is more representative of the olfactive universe of perfumery than that of Arctander, similar results were obtained. These studies confirmed that odor descriptors used in perfumery are generally rather independent, with no strict hierarchy among them, ruling out the existence of a small number of primary odors.

Another detailed database is the Atlas of odor character profiles (Dravnieks 1985) that contains the odor profile of 138 pure odorant chemicals. Data were collected from 120–140 panelists at 12 participant laboratories. A list of 146 commonly used descriptors was provided to the panelists, who smelled the sample and described its odor by rating the applicability of each descriptor on a numeric scale from 0 to 5. In a previous publication (Dravnieks 1982), these average profiles exhibited an impressive reliability. Because the number of chemicals in this database is not large enough for a proper characterization of odor space, additional compounds were profiled using the same descriptors with a panel

of about 20 individuals (Jeltema and Southwick 1986), resulting in a compilation of 415 odorants. Further experiments indicated that the results from this reduced panel correlated well with those from the Dravnieks' panel. This database was analyzed using factor analysis. The identification of descriptors that contributed the most to each factor allowed their classification into 17 groups of terms.

The "Sigma–Aldrich Fine Chemicals (SAFC) flavors and fragrances" catalog is another large database of semantic odor profiles. A recent work (Madany-Mamlouk et al. 2003; Madany-Mamlouk and Martinetz 2004) has compiled 278 descriptors from the 1996 edition of this catalog that comprised 851 perfume raw materials (PRMs). The application of multidimensional scaling (MDS) to this database revealed approximately 32 dimensions in the olfactory perception space, which agrees with the long-held belief that olfactory space is high dimensional. Afterward, a 2-dimensional self-organizing mapping was used to visualize the MDS results on a low-dimensional map. This map provides some sort of clustering for odor descriptors, but those that appear as neighbors might actually be very distant in the high-dimensional space. Given this drawback, a new statistical effort is described in this paper that was conducted to determine if a clearer classification of odor descriptors could be achieved and allow for a better understanding of the underlying structure in human odor perception.

Materials

The SAFC flavors and fragrances catalog 2003–2004 (Sigma-Aldrich 2003), hereafter referred to as "SAFC catalog," was used as the source of the data for this study. An updated version can be requested from the Web (http://www.sigmaaldrich. com/SAFC/Supply_Solutions_Flavors.html). This catalog presents 29 main odor categories under the "organoleptic properties" section. Seven of them are subdivided into a different number of subcategories: fruity (18), citrus (4), floral (15), herbaceous (4), nutty (5), balsamic (8), and fatty (6), resulting in a set of 60 odor classes. Including in this list the 22 "independent" odor categories that appear with no subdivision, it results in a pool of 82 odor classes, listed in Table 2. So, a particular objective of this study was to identify similarities among the 22 independent odor categories. The SAFC catalog contains 881 PRMs, with one or more PRMs listed for each one of the 82 odor categories. Most of PRMs appear under more than one category. These materials will be referred to as PRMs and not just "compounds" because some of them are mixtures. A list of the 881 PRMs can be found at http://www.geocities.com/mazarcas /SAFC_list.htm. According to SAFC Flavors and Fragrances, odor profiles have been basically obtained from the literature (Arctander 1969; Dravnieks 1985; Burdock 2004), supplemented with feedback from industrial customer's flavorists, or through interactions with the Chemical Sources Association (M McNello, personal communication).

The information was organized by collecting the descriptors assigned to each PRM. As reported in the analysis of similar databases (Chastrette et al. 1988), a matrix was created containing 881 PRMs as observations (in rows) and 82 dichotomic variables (in columns), each one representing a particular odor descriptor. In this matrix, the element x_{ij} takes the values 1 if the descriptor j is present in the odor profile of the PRM i and 0 when that note is absent, which is actually more frequently the case. This "dichotomic" matrix, which contains the semantic descriptions coded numerically, is the "SAFC database" with which the statistical analyses have been conducted.

Methods

A descriptive analysis was conducted with 2 variables: number of descriptors assigned to a given PRM and number of PRMs assigned to a given odor descriptor. In order to identify associations between descriptors, the correlation coefficient for all possible pairs of descriptors was computed and those with higher values were identified. Each odor descriptor of the SAFC database is a dichotomic variable with 2 values (0 and 1), so that the sum corresponds to the number of PRMs labeled with that particular descriptor. If the entries of a given descriptor are randomly scrambled, it results in a new random descriptor with the same sum but uncorrelated with the original one. This procedure was applied to the 82 descriptors, resulting in a matrix of independent (orthogonal) variables. This new matrix of random descriptors presents the same size as the original one and was referred to as the "random matrix."

Principal components (PCs) are directions of maximum data variance obtained as linear combinations of the original variables. The projections of observations (PRMs in this case) over these directions are called "scores," and the contributions of the variables (odor descriptors) in the formation of a given component are called "loadings." A scatter plot of the loadings corresponding to 2 different components is referred to as the "loading plot." PCA was used to evaluate both the randomized and original matrices. The comparison of these analyses allowed for the identification of those descriptors to be discarded because they represent too few PRMs to be useful. A new PCA was fitted using the remaining descriptors. An examination was made of the loading plots for the different components in order to explore the odor space of this database and to identify clusters of similar notes. Once a set of descriptors was identified that clearly defined a component or dimension of odor space, these variables were set aside and a new PCA was calculated in order to identify the next dominant latent structure. Using this cascaded PCA approach, a classification of descriptors was finally obtained. All PCAs were carried out using the software SIMCA-P 10.0 (http://www.umetrics .com). The data were centered and scaled to unit variance prior to analysis.

Results and discussion

Descriptive analysis of the SAFC database

The organoleptic properties section of the SAFC catalog presents the compounds under the 82 odor character categories. Another section lists the compounds alphabetically, and most of them contain an unrestricted odor description with a few words chosen from a larger list; many of them (like the unpleasant notes) are not included in the set of 82 descriptors. This additional information has been used in a reported analysis of the 1996 edition of this catalog (Madany-Mamlouk et al. 2003; Madany-Mamlouk and Martinetz 2004), but because this odor profile is not available for all compounds, we have used exclusively the information under the organoleptic properties section.

The number of odor descriptors assigned for a given PRM ranges from 1 to 9, with an average value of 2.2. The occurrences of a given descriptor (number of PRMs labeled with that descriptor) range from 1 to 141, with an average of 24 (Figure 1). In comparison, the Arctander database contains 233 descriptors that provide the relevant olfactory information. Each note was cited an average 29 times, and the average number of words used to describe the odor of a particular compound was 2.7 (Chastrette et al. 1988). Thus, these characteristics are similar in the SAFC database.

Identification of associated odor descriptors

The linear correlation coefficient (r) was calculated for all possible pairs of descriptors, except those with an occurrence of <6 PRMs (too few to provide relevant information). This coefficient is widely used to study the correlation between 2 continuous variables but not often for dichotomic ones as this case. It provides an easy interpretation: $r = 1$ if 2 descriptors are identical and r will approach 0 if there is no similarity or correlation. Thus, r can be used as a measure of similarity. The highest 60 values out of the 3321 possible

pairs are shown in Table 1. In all cases, the correlation is statistically significant (P value < 0.002). The similarity detected between most of these odor character descriptors was intuitively appealing, and this information will be used later to discuss the PCA results.

Other workers have reported using the product of the original dichotomic matrix and its transpose (XX^T) to generate an occurrence/co-occurrence matrix where the diagonal

Table 1 Pairs of descriptors with highest correlation

r	Pair of descriptors	r	Pair of descriptors
0.288	Rose–honey	0.133	Apricot–berry
0.266	Cinnamon–spicy	0.132	Butter–oily
0.265	Pineapple–banana	0.127	Apricot–plum
0.242	Butter–cheese	0.126	Apricot–apple
0.238	Apricot–pineapple	0.126	Musty–fatty-other
0.228	Apricot–peach	0.125	Pineapple–apple
0.213	Vanilla–caramel	0.124	Pear–grape
0.190	Meaty–nutty-other	0.123	Apricot–cherry
0.180	Sulfurous–alliaceous	0.122	Apple–banana
0.178	Cherry–raspberry	0.121	Raspberry–violet
0.175	Butter–creamy	0.121	Winelike–fruity-other
0.172	Minty–camphoraceous	0.120	Peach–butter
0.169	Sulfurous–meaty	0.120	Hazelnut–chocolate
0.168	Plum–cinnamon	0.120	Coconut–creamy
0.166	Waxy–fatty-other	0.117	Cherry–strawberry
0.164	Coffee–alliaceous	0.117	Citrus-other–rose
0.162	Hyacinth–balsam	0.115	Peach–coconut
0.161	Peach–pear	0.114	Coffee–nutty-other
0.159	Citrus-other–waxy	0.111	Peach–berry
0.158	Floral-other–sweet	0.111	Violet–green
0.152	Meaty–alliaceous	0.110	Fruity-other–ethereal
0.151	Apple–plum	0.109	Raspberry–strawberry
0.144	Cherry–plum	0.109	Pineapple–grape
0.144	Coffee–meaty	0.107	Apricot–honey
0.143	Hazelnut–almond	0.106	Vegetable–sulfurous
0.142	Hazelnut–earthy	0.105	Cinnamon–hazelnut
0.142	Vegetable–green	0.105	Cinnamon–hyacinth
0.137	Smoky–woody	0.105	Medicinal–camphoraceous
0.137	Hazelnut–woody	0.104	Cheese–animal
0.134	Almond–sweet	0.104	Musty–woody

Sixty pairs of odor descriptors with the highest linear correlation coefficient (r), sorted in descending order. Descriptors with an occurrence <6 are not included.
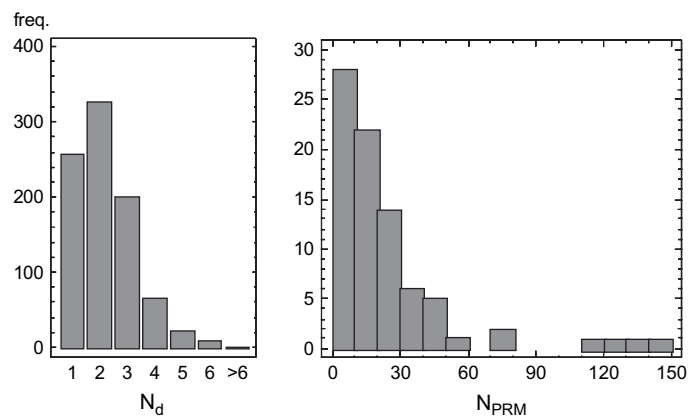

Figure 1 Bar chart for the number of odor descriptors assigned to a given PRM (N_d) (Left). Histogram of the number of PRMs described with a given descriptor (N_{PRM}) (Right). The vertical scale corresponds to absolute frequency.

terms are the occurrences of notes and nondiagonal terms represent the co-occurrences between notes (Chastrette et al. 1986, 1991). This matrix is then transformed into a similarity matrix, normally used for the multivariate analysis. In a reported study of the Firmenich database (Chastrette et al. 1991), 4 different multivariate methods were applied using this type of similarity matrix, and PCA was found to be less suitable for analyzing the relationships among descriptors than the other methods. However, PCA is a useful tool to study the underlying latent variables in a matrix structured in observations (PRMs) by variables (odor descriptors), which is not the case with the similarity matrix. Thus, we wondered if the analysis of the original dichotomic matrix with PCA would lead to more interpretable results.

Identification of descriptors that do not provide relevant information

In the analysis of the Arctander database (Chastrette et al. 1988), 2467 compounds were selected and combined descriptors were used to replace pairs of very similar notes (amber/ambergris, citrus/lemon, and raspberry/berry). Next, 37 descriptors were eliminated for being considered either rather subjective or related with intensity, such as "dry," "fresh," "strong," "weak," "warm," or "deep." Lastly, odor descriptors with fewer than 12 occurrences were discarded (156 in total). In our case, we skipped both of these steps in order to let the analysis identify notes with high similarity. If we were to perform the same analysis for the SAFC database of 881 PRMs, we would need to discard descriptors yielding 4.3 or fewer occurrences (maintaining the proportion: $4.3 = 12 \times 881/2467$).

To check if this criterion is adequate, a PCA was conducted with the random matrix, obtained by randomly scrambling the entries of the original matrix, as described above. This analysis showed that the descriptors "soapy," "mossy," "pepper," "lime," and "gardenia" (with 4, 4, 2, 1, and 1 occurrences, respectively) were forcing the first 2 PCs. Discarding the 20 descriptors with <5 occurrences and repeating the PCA, those with 5 or 6 occurrences did not force components. This criterion coincides with the one previously observed.

For this PCA with the random matrix and for an equivalent PCA with the original matrix (62 descriptors), the eigenvalues and percentage of data variance explained by each component (goodness of fit, R_x^2) have been compared (Figure 2). A weak correlation structure is clearly observed in the dichotomic matrix, with each component explaining just slightly more of the variance than the random case. This suggests that odor descriptors are quite independent and the correlation between most descriptors is too weak to define a PC.

Regarding the data pretreatment more convenient for PCA, 3 choices are possible: centered variables, scaled to unit variance, or both. In this case, the average and variance of an odor descriptor increase according to the number of occurrences. If a PCA is fitted with the original data (no pretreat-
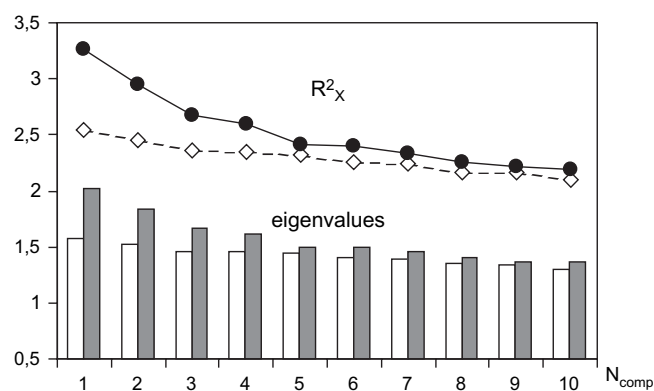

Figure 2 Characteristics of the PCs from the odor profile dichotomic matrix (62 descriptors): R_x^2 (filled symbols) and eigenvalue (filled bars). Characteristics of the PCs from the random matrix (with the same structure as the original dichotomic matrix but with entries of each column randomly scrambled): R_x^2 (open symbols) and eigenvalue (open bars). Vertical scale: R_x^2 in percentage and eigenvalues in nondimensional units.

ment), the components are forced by the descriptors appearing most frequently. However, this observation is not related to the similarities between descriptors. For this reason, all PCA models have been fitted with data centered and scaled to unit variance.

Identification of clusters: noncitrus fruity

Conducting a PCA with the 62 descriptors assigned to at least 5 PRMs and checking the score plot for PC1 and PC2 (projection of PRMs over the 2 PCs), 2 orthogonal directions of variability appear (Figure 3). The corresponding loading plot reveals that PC1 corresponds to the noncitrus fruity descriptors. Thus, the strongest dimension of the SAFC database is defined by the fruity odorants. The highest loadings in absolute value along this direction correspond to the descriptors that best characterize the fruity odor, and the most representative is "apricot." This is the fruity attribute most frequent in Table 1. Highest loadings and proximity in the loading plot correspond to correlated descriptors that are used by panelists interchangeably. This is the case for "pineapple–banana," the third pair with the highest correlation (Table 1). The fact that the notes "raspberry," "strawberry," "grape," and "melon" are closer to the center might indicate that these notes are less characteristically fruity. However, this conclusion is uncertain, given that the number of PRMs labeled with these descriptors is lower compared with the rest of descriptors in the fruity cluster (Table 2). "Coconut" appears in the SAFC catalog under the fruity category, but according to Figure 3, it is the only noncitrus fruit excluded from the cluster. Other authors have considered that coconut odor is related with nuts (Jeltema and Southwick 1986; Chastrette et al. 1988; Madany-Mamlouk et al. 2003). However, the highest correlation of this descriptor corresponds to creamy and peach (Table 1). Thus, it was classified as intermediate of fruity and butter (Table 2).
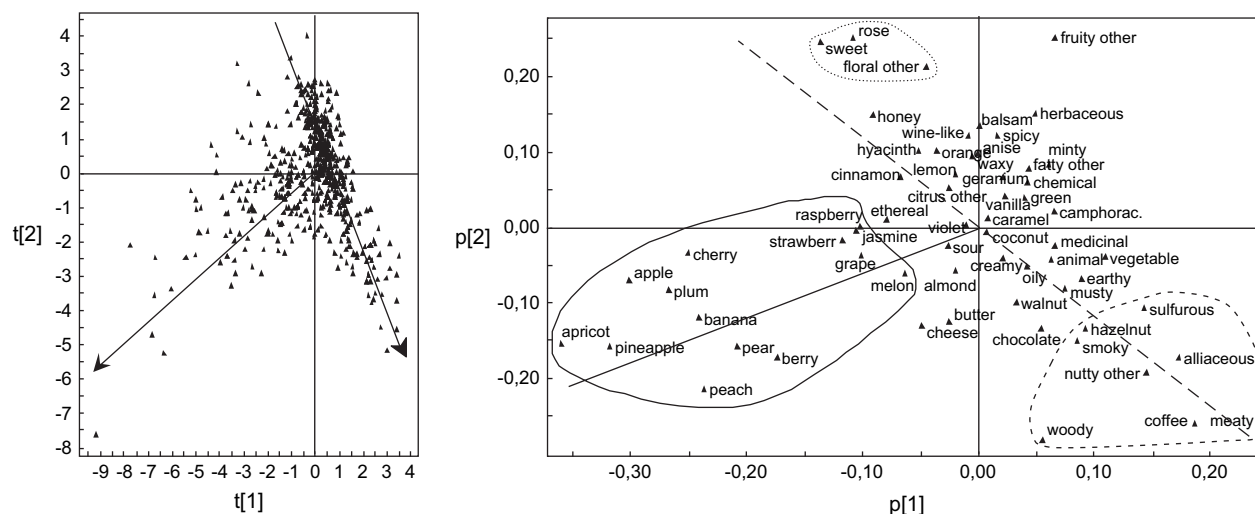
Figure 3 Results of the PCA with 62 relevant descriptors (occurrence > 4). Score plot (left) and loading plot (right) corresponding to the first 2 PCs (PC 1 and PC 2).

Strikingly, the descriptor "fruity-other" is far separated from the fruity cluster. This note is assigned to 131 PRMs, and only 4 of them are also labeled with another noncitrus fruity note. Thus, this descriptor was used only when the PRM smells fruity, but not like any one in particular, which appears as a negative correlation in the PCA: if "fruity-other" = 1, then the rest of fruity notes are more likely to be 0. As a consequence, the PCA reflects no similarity between "fruity-other" and the rest of fruity descriptors, but obviously, this note should be included in the noncitrus fruity cluster. "Ethereal" is correlated with "fruity-other" (Table 1) and appears close to the fruity cluster. Actually, ethereal and fruity are close odors, according to the experience of perfumers (Chastrette et al. 1988, 1991), and one of the odor classification schemes proposes the class ethereal instead of fruity (Zwaardemaker 1925). Nevertheless, in certain studies, the "etherish" note was reported similar to "chemical" or "medicinal" but not to "fruity" (Jeltema and Southwick 1986).

Different studies have reported that when a wide range of odors are sampled, a common result is a configuration of odor space with one underlying hedonic dimension, related with the pleasant–unpleasant degree of odor perception (Woskow 1968; Davis 1979). In this case, the hedonic dimension does not appear clearly, probably because the database is not representative of odor perception space, being biased toward those odors most frequent in perfumery that are in general rather pleasant. However, the loading plot PC1–2 (Figure 3) suggests that the second component might be related with the hedonic dimension. So, if the descriptors are orthogonally projected over the dashed line, the most pleasant notes tend to be located in one extreme, whereas the most unpleasant seem to appear on the other extreme (dashed cluster).

Identification of clusters: butter and alliaceous

Because noncitrus fruity descriptors are clearly grouped, these notes used to form the cluster are set aside in order

to identify which other descriptors define a clear direction of variability. The analysis was conducted using 48 descriptors remaining after the removal of the noncitrus fruity descriptors: "apricot," "pineapple," "apple," "plum," "cherry," "banana," "pear," "peach," "berry," "strawberry," "raspberry," "grape," "melon," and "fruity-other."

The loading plot PC1–2 of the previous model (Figure 3) reveals that the components are rotated. Moreover, the loadings of PC2 are scattered, with no clear distinction of clusters. This situation is common in PCA. Consequently, instead of using automatic procedures for cluster identification, it is preferable to rely on visual inspection methods, checking loading plots with different combinations of components in order to find which plot reveals some descriptors clustered together and clearly separated from the rest. The loading plots PC1–2 and PC3–4 usually provide the most relevant information, given that the first components account for the highest data variability. But this is not necessarily the case in this PCA with 48 descriptors because there are no clear dominant components (the different PCs explain a similar amount of the total data variance). The loading plot for PC3–5 (Figure 4) reveals that PC3 is dominated by the notes "butter," "cheese," and "creamy." A significant similarity between "buttery" and "creamy" was also identified in the Arctander database (Chastrette et al. 1988). In the other reported study of the SAFC catalog (Madany-Mamlouk et al. 2003), a cluster was formed with the notes "butter," "creamy," and "milk." "Oily" has been classified as an intermediate odor between fatty and butter. "Coconut" is also close to this cluster, revealing the buttery smell of this note. The presence of "vanilla" and "caramel" not far from this cluster reveals that butter, creamy, and cheese are pleasant smells with a certain similarity to balsamic notes.

The descriptors "alliaceous" (garlic, onion smell) and "sulfurous" form an independent dimension, revealed by PC5. Other authors have also proposed this cluster as an

Table 2   Proposed classification of odor descriptors

Descriptor	N_{PRM}	SAFC classification	Proposed classification
Apricot	23	Fruity	Fruity
Pineapple	40	Fruity	Fruity
Apple	44	Fruity	Fruity
Peach	18	Fruity	Fruity
Banana	24	Fruity	Fruity
Plum	18	Fruity	Fruity
Cherry	19	Fruity	Fruity
Pear	16	Fruity	Fruity
Berry	28	Fruity	Fruity
Strawberry	12	Fruity	Fruity
Grape	13	Fruity	Fruity
Raspberry	6	Fruity	Fruity
Melon	8	Fruity	Fruity
Fruity-other	131	Fruity	Fruity
Coconut	13	Fruity	Fruity, butter
Quince	1	Fruity	(Fruity)
Jam	4	Fruity	?
Grapefruit	2	Fruity	?
Orange	13	Citrus	Citrus
Lemon	9	Citrus	Citrus
Citrus-other	36	Citrus	Citrus
Lime	1	Citrus	(Citrus)
Rose	53	Floral	Floral
Jasmine	18	Floral	Floral
Hyacinth	11	Floral	Floral
Geranium	9	Floral	Floral
Floral-other	74	Floral	Floral
Violet	10	Floral	Floral, green
Lily	4	Floral	(Floral)
Gardenia	4	Floral	(Floral)
Blossom	4	Floral	(Floral)
Carnation	3	Floral	(Floral)
Lilac	2	Floral	(Floral)
Narcissus	1	Floral	(Floral)
Marigold	1	Floral	(Floral)
Jonquil	1	Floral	(Floral)
Iris	1	Floral	(Floral)
Herbaceous-other	75	Herbaceous	Herbaceous
Sage	3	Herbaceous	(Herbaceous)
Caraway	2	Herbaceous	(Herbaceous)
Clove	2	Herbaceous	?
Hazelnut	11	Nutty	Nutty
Walnut	5	Nutty	Nutty
Nutty-other	33	Nutty	Nutty
Almond	15	Nutty	Nutty

Table 2   Continued

Descriptor	N_{PRM}	SAFC classification	Proposed classification
Peanut	2	Nutty	(Nutty)
Balsam	26	Balsamic	Balsamic
Vanilla	17	Balsamic	Balsamic
Chocolate	20	Balsamic	Balsamic
Caramel	18	Balsamic	Balsamic
Cinnamon	7	Balsamic	Balsamic
Anise	12	Balsamic	Balsamic
Honey	23	Balsamic	Balsamic, floral
Sweet	141	Balsamic	Balsamic, floral
Spicy	46	Independent	Balsamic
Butter	25	Fatty	Butter
Creamy	17	Fatty	Butter
Cheese	21	Fatty	Butter
Fatty-other	44	Fatty	Fatty
Oily	38	Fatty	Fatty, butter
Sour	5	Fatty	Fatty, fruity
Green	124	Independent	Green
Vegetable	19	Independent	Green
Earthy	32	Independent	Wet
Musty	23	Independent	Wet
Mossy	4	Independent	(Wet)
Alliaceous	29	Independent	Alliaceous
Sulfurous	14	Independent	Alliaceous
Camphoraceous	21	Independent	Camphoraceous
Minty	31	Independent	Camphoraceous
Medicinal	26	Independent	Camphoraceous, chemical
Chemical	12	Independent	Chemical
Winelike	25	Independent	Chemical, fruity
Ethereal	48	Independent	Chemical, fruity
Meaty	48	Independent	Cooked
Coffee	24	Independent	Cooked
Smoky	14	Independent	Cooked
Woody	117	Independent	Woody
Animal	13	Independent	Animal
Waxy	28	Independent	Waxy
Pepper	2	Independent	?
Soapy	4	Independent	?

For each one of the 82 odor character descriptors of the SAFC catalog, the table presents the number of PRMs labeled with that descriptor (N_{PRM}), the classification assigned in the SAFC catalog, and the proposed classification according to the results of PCA (in some cases, the descriptor has been considered as intermediate of 2 odor classes). Descriptors in parentheses were not considered in the analysis ($N_{PRM} < 5$). The term "independent" indicates that the descriptor appears in the SAFC catalog as an independent or isolated odor class. The question mark indicates those descriptors with an uncertain classification.
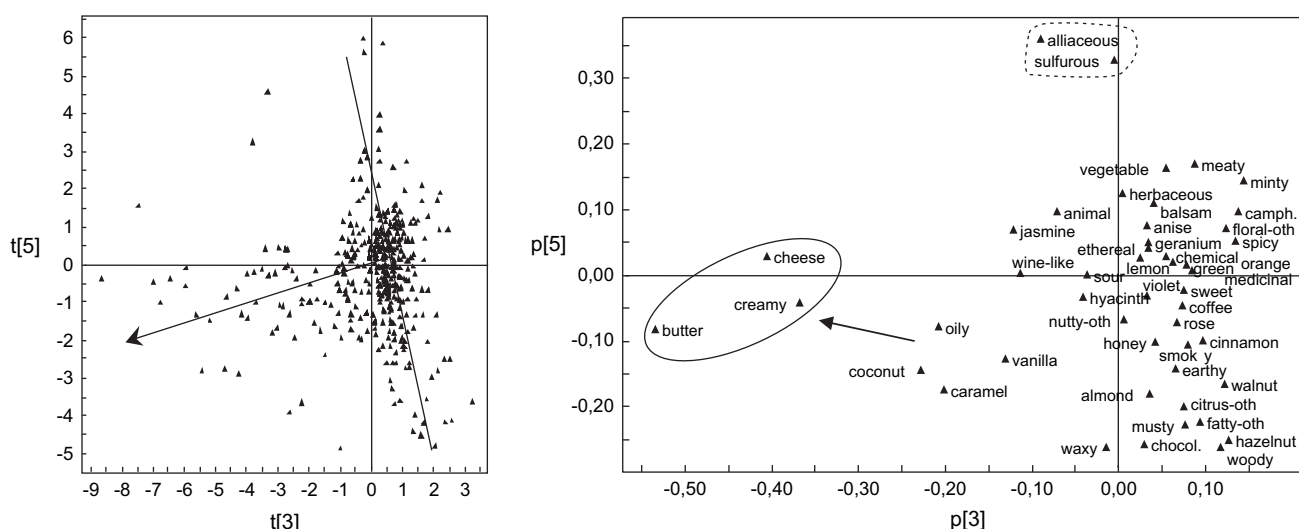
Figure 4 Results of the PCA with 48 descriptors (model of Figure 3 discarding the noncitrus fruity descriptors). Score plot (left) and loading plot (right) for PC3 and PC5.

odor class (Zwaardemaker 1925; Jeltema and Southwick 1986).

Identification of clusters: balsamic, nutty, and camphoraceous

As before, the next analysis was conducted using the odor descriptors not already accounted for in previous clusters. Thus, a new PCA was fitted using the 43 descriptors remaining after discarding from the previous model the descriptors "butter," "cheese," "creamy," "alliaceous," and "sulfurous." The loading plot PC1–2 (Figure 5) reveals that the second component is related with balsamic descriptors. The SAFC catalog considers as balsamic notes: "vanilla," "sweet," "honey," "cinnamon," "chocolate," "caramel," "balsam," and "anise." Most of them correspond to the highest loadings in PC2, suggesting that they are related odors, but they do not form a compact cluster. Thus, their classification is discussed in the next model.

In this PCA with 43 descriptors, the first component is related to notes with a rather different smell: "woody," "meaty," "coffee," "smoky," and "nutty." Checking different loading plots (PC1–2, PC1–3, PC1–4, PC2–3, etc.), the one for PC1–4 (Figure 5) reveals that the nutty descriptors are located close to each other, separated from "meaty," "coffee," and "smoky". This cluster appears close to "cinnamon," "earthy," and "woody," 3 notes with a certain similarity to nutty (Table 1). The descriptor "nutty-other" is close to "hazelnut," "walnut," and "almond" in the loading plot PC1–2, but this proximity is not reflected in the loading plot PC1–4. The reason for this is likely to be similar to the observation regarding the "fruity-other" descriptor in the case of the noncitrus fruity cluster. Other authors have also proposed an independent nutty category (Jeltema and Southwick 1986).

In the loading plot for PC3–5, a cluster can be clearly observed that comprises "minty," "camphoraceous," and

"medicinal," and the score plot indicates a set of PRMs following this direction. This cluster has also been proposed in other studies (Jeltema and Southwick 1986). In the analysis of the Firmenich database (Chastrette et al. 1991), "minty" appeared related to "hay," whereas "camphoraceous" was similar to "piney" and not distant from "woody." A close look at this loading plot reveals that although the 3 clustered descriptors are separated from the rest, "minty–herbaceous" are located not too far away, and the same occurs for "camphoraceous–woody" and "medicinal–chemical." Other works have also found a similarity between "camphor" and "minty" (Chastrette et al. 1988; Madany-Mamlouk et al. 2003) but not so with "medicinal," reported to be found similar to "chemical," "etherish" (Jeltema and Southwick 1986), and "phenolic" (Chastrette et al. 1988; Abe et al. 1990).

From the previous model with 43 variables, a new PCA was conducted by first eliminating the notes "minty," "camphoraceous," "almond," "walnut," "hazelnut," and "nutty-other." "Medicinal" was also included to check other possible similarities. The loading plot PC1–2 (Figure 6) reveals that the notes in the upper part of the plot correspond to "spicy" and the 8 descriptors classified as balsamic in the SAFC catalog except "honey" and "anise." These results complement the previous PCA (Figure 5), highlighting that balsamic notes form an independent dimension in odor space.

"Honey" and "rose" are close in the loading plots, and they present the highest correlation of all pairs of descriptors (Table 1). This high similarity has also been pointed out in other studies (Chastrette et al. 1988, 1991). Thus, "honey" has been classified as a note intermediate between balsamic and floral.

The "spicy" descriptor seems controversial. A significant similarity between "spicy," "herbaceous," and "aromatic" was identified in the Arctander database (Chastrette et al. 1988). In a reported study using the Dravnieks' descriptors, "spicy"
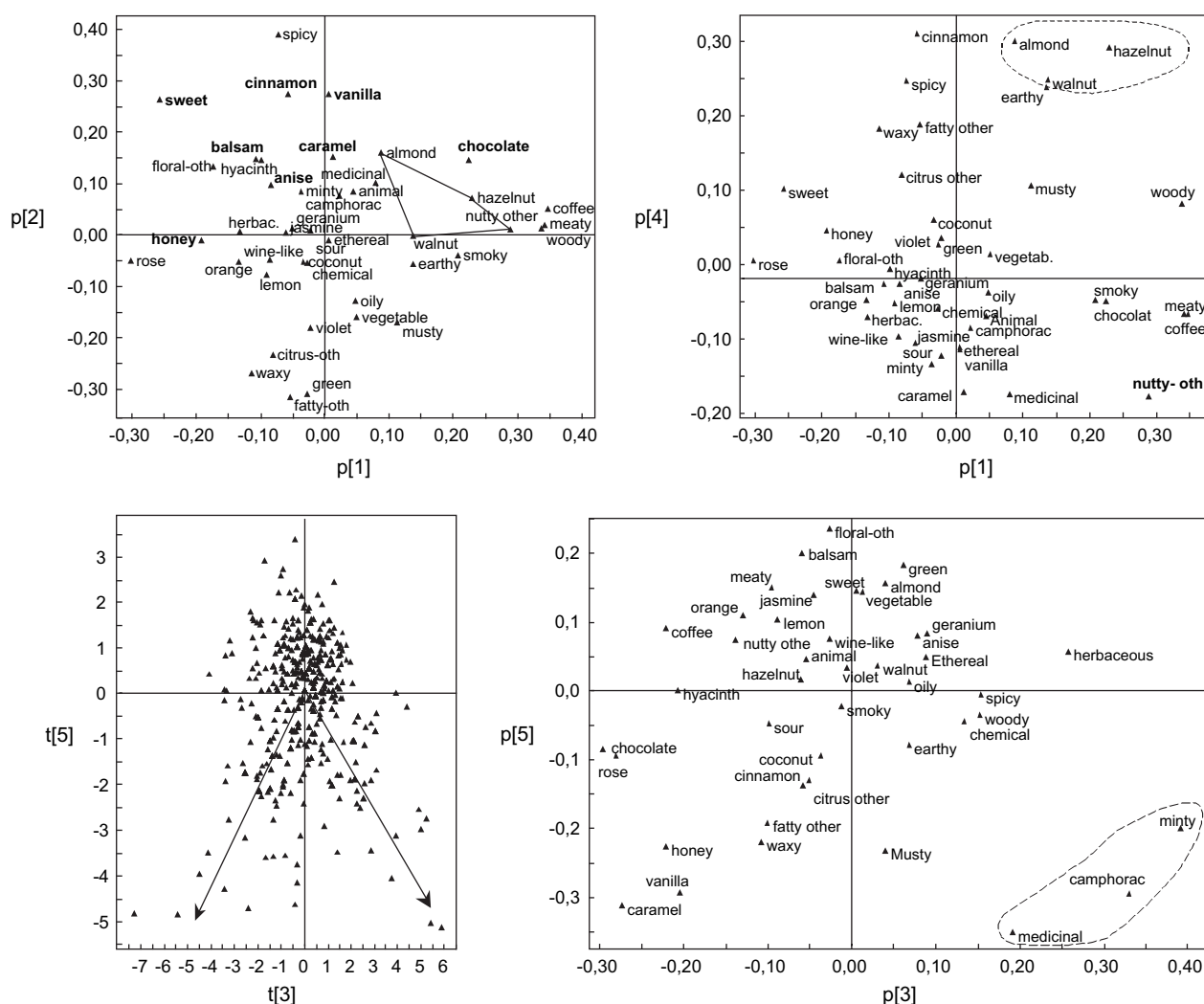
Figure 5 Results of the PCA with 43 descriptors (model of Figure 4 discarding the butter and alliaceous clusters). Loading plot for PC1–2 (upper left), PC1–4 (upper right), and PC3–5 (lower right). Score plot for PC3–5 (lower left).

was included with the cinnamon group but not with "vanilla," "chocolate," "caramel," or "honey" (Jeltema and Southwick 1986). In the SAFC catalog, "spicy" is not included within the balsamic category. But according to our results, the correlation coefficient for "spicy–cinnamon" is the second highest of all pairs of descriptors (Table 1). Figures 5 and 6 show that "spicy" is strongly associated with the balsamic notes. This result agrees with the reported analysis of the Firmenich database (Chastrette et al. 1991), where a significant similarity between "spicy" and "balsamic" was found.

Another descriptor with a troublesome classification is "sweet." According to our results, "sweet" can clearly be considered within the balsamic cluster but remains close to the "floral-other" note (Figures 5 and 6) because the highest correlation of "sweet" corresponds to "floral-other" and "almond" (Table 1). A reported analysis of the SAFC catalog (Madany-Mamlouk et al. 2003) also showed the "sweet" classifier to be related to the descriptors "pleasant"

and "spicy." But other authors (Jeltema and Southwick 1986) classify this note with the noncitrus fruits, not with other balsamics. In this analysis, we have decided to classify it as intermediate between balsamic and floral.

"Anise" is listed under the balsamic category in the SAFC catalog, but in Figures 5 and 6, it is the balsamic note closest to the center of the loading plot and hence can be considered as the least balsamic. Other authors have also classified "anise" in a group different to other balsamic notes (Zwaardemaker 1925; Jeltema and Southwick 1986).

Identification of cluster: cooked

Discarding from the previous model "spicy," "coconut," and the 8 balsamic descriptors and conducting a new PCA with the remaining 27 variables, a cluster was identified in the loading plot PC1–2. The new cluster comprises the notes "woody," "meaty," "coffee," and "smoky" (figure
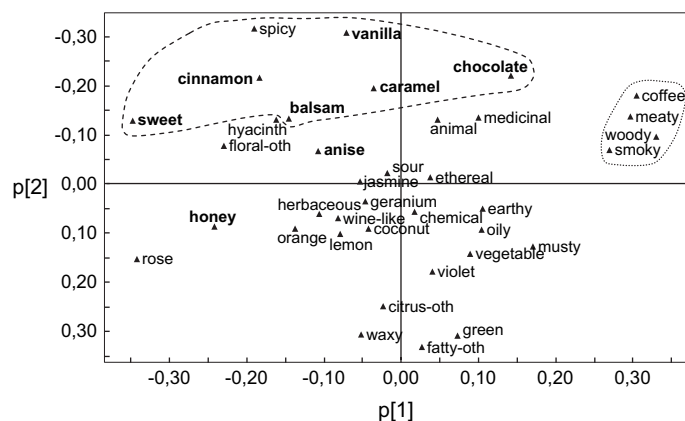
Figure 6 Results of the PCA with 37 descriptors (model of Figure 5 discarding the nutty and camphoraceous clusters). Loading plot for PC1–2. Descriptors highlighted in bold are those under the balsamic category in the SAFC catalog. Values of p[2] are shown in reverse order to ease the comparison with Figure 5.

not shown). This cluster appears more clearly with PC1–5 (Figure 7) and is also apparent in Figure 6.

The relationship "coffee–smoky" is intuitively appealing. Checking the correlation with the rest of descriptors (Table 1), it appears that this cluster is related with the nutty and the alliaceous odors. Different works have found similarities between "woody" and other descriptors: "cognac" (Jeltema and Southwick 1986), "animal" (Madany-Mamlouk et al. 2003), or "amber" (Chastrette et al. 1991) but not with "meaty" or "coffee." According to Table 1, the highest correlation of "woody" corresponds to "smoky," "hazelnut," and "musty." There is probably a subgroup of PRMs with a smoky note among the 117 ones labeled as "woody," and for this reason, it appears close to "smoky" in Figure 7. Given that "woody" sometimes is considered as an independent odor class (Jeltema and Southwick 1986), we have regarded "woody" as an isolated descriptor. "Meaty," "coffee," and "smoky" are close to the nutty cluster (Figure 5), and some authors have included "meaty" and "smoky" within the nutty category (Jeltema and Southwick 1986; Madany-Mamlouk et al. 2003). However, we consider it more appropriate to create a new category that is referred in Table 2 as cooked.

Identification of clusters: floral, citrus, and green

As before, a new analysis was conducted by discarding these 4 descriptors ("woody," "meaty," "coffee," and "smoky") and fitting a new PCA using the remaining 23 variables. The first component of the new model is dominated by "waxy" and "fatty-other." The descriptors shown in the loading plot PC2–3 (Figure 8) above the dashed line correspond to the categories floral, citrus, and green. The proximity between "floral," "rose," "citrus," "herbaceous," and "green" was also found in other studies (Chastrette et al. 1991). "Rose" presents the highest negative loadings in PC2 and hence is the most representative of floral notes.
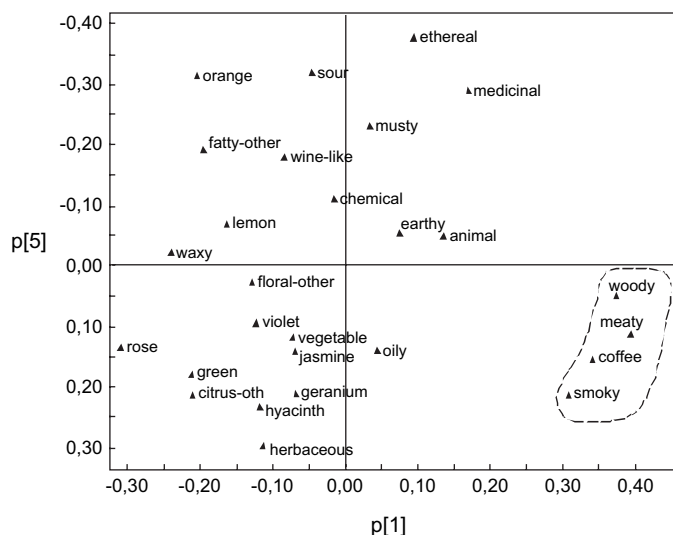

Figure 7 Results of the PCA with 27 descriptors (model of Figure 6 discarding the balsamic cluster). Loading plot for PC1–5.

Regarding citrus fruits ("orange" and "lemon"), the results (Figures 3 and 8) show that their smell resembles more the floral notes than the rest of noncitrus fruits. Consequently, they have been classified in an independent cluster (Table 2), as reported in other cases (Jeltema and Southwick 1986). In the analysis of the Firmenich database, citrus was found to be more similar to herbaceous and floral than to fruity. The descriptor "citrus-other" appears separated for the same reason as "fruity-other" and "nutty-other."

"Green" refers to the smell of fresh-cut grass, whereas "vegetable" refers to fresh vegetables like green pepper, cucumber, or green beans. Although both appear as independent odor categories in the SAFC catalog, our results (Table 1, Figure 8) reveal that they are related odors, and we grouped them as a single odor class (Table 2). Because vegetables can be described botanically as herbaceous plants (nonwoody annual), one would expect to observe more similarity between the descriptors "green," "vegetable," and "herbaceous." But the term "herbs" is commonly assigned to plants or plant parts used for medicinal, flavoring, or aromatic purposes. In the loading plot PC2–3 (Figure 8), "herbaceous" is clearly separated from "green" and "vegetable" but closer to floral, and a similar result has been reported in other studies (Madany-Mamlouk et al. 2003). Thus, in the SAFC catalog, the note "herbaceous" is used in this second meaning. The same meaning is probably assigned to "herbaceous" in the Arctander database because a significant similarity between "green" and "floral" was found, and also for "herbaceous–spicy" (Chastrette et al. 1988). But in the Dravnieks Atlas, "herbal–green–cut grass" is a single descriptor, according to the botanical definition of herb. Thus, herbaceous was regarded as an independent odor class, following the criterion of the SAFC catalog. "Violet" is the floral note most close to green because it
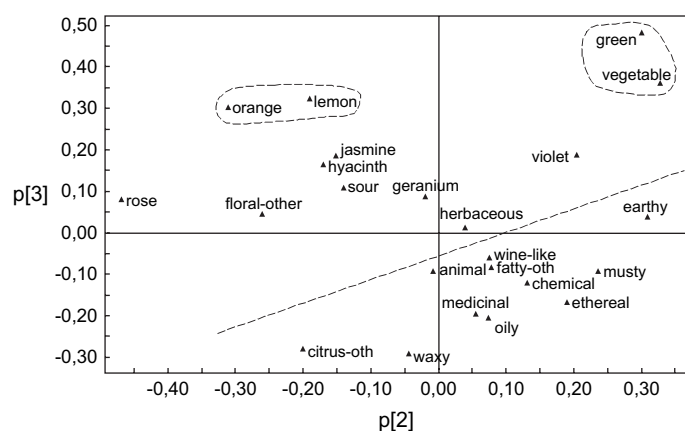
Figure 8   Results of the PCA with 23 descriptors (model of Figure 7 discarding "woody," "meaty," "coffee," and "smoky"). Loading plot for PC2–3. The dashed line separates the floral–citrus–green descriptors from the rest.

appears between "vegetable" and the floral notes (Figure 8). Some authors have found a similarity between "violet" and "leafy" (Chastrette et al. 1988). So, it was regarded as intermediate of floral and green.

Identification of the remaining clusters

At this point, discarding the floral–citrus–green notes would leave only 11 descriptors. This number is too small, and the results do not highlight clear similarities between descriptors. So, the remaining notes have been classified according to information from the literature. Some authors have considered that "musty," "earthy," and "moldy" can be included within the green–vegetable category (Jeltema and Southwick 1986). But according to Figure 8, the odor appears clearly different and a cluster referred to as "wet" has been formed with these descriptors. Actually, one of the proposed odor classification systems (Klein 1947) considers earthy-fungoid as one of the 8 odor classes.

In a reported study, solvent-related descriptors were considered as an independent class, comprising notes like "chemical," "etherish," and "medicinal" (Jeltema and Southwick 1986). In a similar way, we have created a cluster with the "chemical" note; "medicinal" has been considered between camphoraceous and chemical, and "ethereal" between fruity and chemical. The classification of "winelike" is uncertain. There are 25 PRMs labeled with this descriptor and also with other notes like "fruity" (14), "sweet" (5), "green" (5), "fatty" (5), or "ethereal" (4). The most reasonable classification is fruity–alcoholic, and consequently, we have regarded it as intermediate of fruity and chemical (Table 2).

"Fatty-other" has been considered as part of a fatty cluster, classifying "oily" as intermediate between fatty and butter. The "sour" note appears in the SAFC catalog under the fatty category, and some authors have classified "sour" with other descriptors like "oily," "fatty," and "cheese" (Jeltema and Southwick 1986). But there are only 5 PRMs labeled as sour, and none of them are additionally described with any

other fatty-related descriptor. On the contrary, 2 of them are also labeled as fruity and 1 as orange, which makes sense given that green fruits are perceived as sour. Thus, we decided to classify "sour" as intermediate of fatty and fruity. A reported study of the Arctander database classified "waxy" in the fatty cluster (Abe et al. 1990). In the SAFC catalog, there are 28 PRMs described as waxy, and most of them are also labeled with different descriptors: "fatty-other" (7), "fruity" (8), "floral" (7), "sweet" (6), "citrus-other" (6), etc. These are rather different odors, and consequently, we decided to classify "waxy" as an independent odor class, following the criteria of the SAFC catalog.

Although some works have classified the "animal" note with other very different odors like "putrid," "fecal," or "oily" (Jeltema and Southwick 1986), it is usually considered associated with "musk" (Chastrette et al. 1988, 1991). In perfumery, musk defines an independent odor category, and amber and musk have long been considered as "ambrosiac" (Zwaardemaker 1925). So, because "musk" does not appear explicitly in the SAFC database, we have considered "animal" as an isolated note (Table 2). This is also justified by the fact that "animal" has a low correlation with the rest of the descriptors, appearing nearly in the last place in Table 1.

Classification of odor descriptors

Regarding the 20 descriptors with an occurrence <5 that were not included in the multivariate analysis, the classification of 5 of them remains uncertain ("jam," "grapefruit," "clove," "pepper," and "soapy"), and the rest were classified like those descriptors with a similar source of the smell according to the criteria of the SAFC catalog: floral ("lily," "gardenia," "blossom," "carnation," "lilac," "narcissus," "marigold," "jonquil," and "iris"), herbaceous ("sage" and "caraway"), fruity ("quince"), citrus ("lime"), and nutty ("peanut").

With the information gathered from all PCAs, 74 odor descriptors have been regrouped in 14 odor classes (some of them as intermediate of 2 categories), and 3 descriptors were considered as independent odors (waxy, woody, and animal), as shown in Table 2. This analysis of odor perception space is restricted by the available descriptors in the SAFC database. Obviously, other descriptors not explicitly included may also form additional odor dimensions.

Conclusions

The results of our statistical analysis of the SAFC database appear to be consistent with the long-held theory that odor space is highly multidimensional. The results suggest that it is reasonable to classify odor descriptors in >9 classes, contrary to many odor classification systems proposed several decades ago and in accordance to more recent statistical analyses of odor profile databases. Understanding the similarities between descriptors and their classification will be helpful in training sensory panels for odor profiling and in providing a standard means of communication among perfumers.

Moreover, this information is also of interest in SOR studies, which can be of considerable value in elucidating the mechanisms of olfaction. These studies are usually focused on particular descriptors, but a different approach would be to use the latent variables of odor space. So, specific SORs for "apricot" or "apple" might be difficult to obtain, but given that all fruity descriptors are related odors, as revealed in this study, it is more reasonable to start with SORs for the whole fruity category and, once the molecular features responsible for this odor class were identified, proceed afterward with a particular fruity odor. Similarly, it seems reasonable to derive SORs for the combined sulfurous–alliaceous category and try next to discriminate between both odors.

Acknowledgements

References

Abe H, Kanaya S, Komukai T, Takahashi Y, Sasaki S. 1990. Systematization of semantic descriptions of odors. Anal Chim Acta 239:73–85.

Amoore JE. 1962. The stereochemical theory of olfaction. I. Identification of seven primary odors. Proc Sci Sect Toilet Goods Assoc 37(Suppl):1–13.

Arctander S. 1969. Perfume and flavor chemicals (aroma chemicals). Volumes 1 and 2. Montclair, NJ: S. Arctander publisher.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–87.

Burdock GA, editor. 2004. Fenaroli's handbook of flavor ingredients. 5th ed. Boca Raton, FL: CRC Press.

Cerbelaud R. 1951. Formulaire de Parfumerie. Paris: Opera.

Chastrette M, de Saint Laumer JY, Sauvegrain P. 1991. Analysis of a system of description of odors by means of four different multivariate statistical methods. Chem Senses 16:81–93.

Chastrette M, Elmouaffek A, Sauvegrain P. 1988. A multidimensional statistical study of similarities between 74 notes used in perfumery. Chem Senses 13:295–305.

Chastrette M, Elmouaffek A, Zakarya D. 1986. Etude statistique multidimensionnelle des similarités entre 24 notes utilisées en parfumerie. C R Acad Sci Ser II Paris 303:1209–14.

Crocker EC, Henderson LF. 1927. Analysis and classification of odors: an effort to develop a workable method. Am Perf Essent Oil Rev 22:325–56.

Davis RG. 1979. Olfactory perceptual space models compared by quantitative methods. Chem Senses 4:21–33.

Dravnieks A. 1982. Odor quality: semantically generated multidimensional profiles are stable. Science 218:799–801.

Dravnieks A. 1985. Atlas of odor character profiles, data series DS 61. Philadelphia, PA: American Society for Testing and Materials.

Dravnieks A, Bock FC, Powers JJ, Tibbetts M, Ford M. 1978. Comparison of odors directly and through profiling. Chem Senses 3:191–225.

Harper R, Bate Smith EC, Land DG. 1968. Odor description and odor classification: a multidisciplinary examination. New York: Elsevier.

Henning H. 1916. Der Geruch. Leipzig, Germany: Barth.

Higuchi T, Shoji K, Hatayama T. 2004. Multidimensional scaling of fragrances: a comparison between the verbal and non-verbal methods of classifying fragrances. Jpn Psychol Res 46:10–9.

Jaubert JN, Gordon G, Doré JC. 1986. Classification of odors and their sensorial perception. Quintessenza 5:27–42.

Jaubert JN, Gordon G, Doré JC. 1987. Une organisation du champ des odeurs. II. Modèle descriptif de l'organisation de l'espace odorant. Parfum Cosmét Arômes 78:71–82.

Jeltema MA, Southwick EW. 1986. Evaluations and application of odor profiling. J Sens Stud 1:123–36.

Kastner D. 1973. Die Beschreibung und klassifizierung von gerüchen. Parfüm Kosmet 54:97–106.

Klein S. 1947. Primary odor element classification. Am Perf Essent Oil Rev 50:453–4.

Lawless HT. 1988. Odor description and odor classification revisited. In: Thomson DMH, editor. Food acceptability. London: Elsevier. p 27–40.

Lawless HT. 1993. Characterization of odor quality through sorting and multidimensional scaling. In: Manley CH, Ho CT, editors. Flavor measurement. New York: Dekker. p 159–83.

Linnaeus C. 1756. Odores medicamentorum. In: Amoenitates Academicae. Volume 3. Stockholm, Sweden: Lars Salvius. p 183–201.

Lovell JH. 1923. Classification of flower odors. Am Bee J 63:392–4.

Madany-Mamlouk A, Chee-Ruiter C, Hofmann UG, Bower JM. 2003. Quantifying olfactory perception: mapping olfactory perception space by using multidimensional scaling and self-organizing maps. Neurocomputing 52–54:591–7.

Madany-Mamlouk A, Martinetz T. 2004. On the dimensions of the olfactory perception space. Neurocomputing 58–60:1019–25.

Malnic B, Hirono J, Sato T, Buck LB. 1999. Combinatorial receptor codes for odors. Cell 96:713–23.

Moskowitz HR, Barbe CD. 1977. Profiling of odor components and their mixtures. Sens Processes 1:212–26.

Rossiter KJ. 1996. Structure–odor relationships. Chem Rev 96:3201–40.

Schiffman SS. 1981. Characterization of odor quality utilizing multidimensional scaling techniques. In: Moskowitz HR, Warren CB, editors. Odor quality and chemical structure. Washington, DC: American Chemical Society. p 1–19.

Schutz HG. 1964. A matching-standards method for characterizing odor qualities. Ann N Y Acad Sci 116:517–26.

Sigma-Aldrich. 2003. Flavors and fragrances 2003–2004 catalog. Milwaukee, WI: Sigma-Aldrich Fine Chemicals Company.

Turin L, Yoshii F. 2003. Structure-odor relations: a modern perspective. In: Doty RL, editor. Handbood of olfaction and gustation. 2nd ed. New York: Marcel Dekker.

Woskow MH. 1968. Multidimensional scaling of odors. In: Tanyolac N, editor. Theories of odors and odor measurement. Bebek, Turkey: Robert College Research Center. p 147–88.

Yoshida M. 1975. Psychometric classification of odors. Chem Senses 1:443–64.

Zwaardemaker H. 1925. L'Odorant. Paris: Doin.