ELSEVIER

0308-8146(95)00087-9

Food Chemistry, Vol. 55, No. 3, pp. 281-287, 1996

0308-8146/96 $15.00+.00
Analytical, Nutritional and Clinical Methods Section

Chemometric classification of honeys according to their type based on quality control data

B. López,[a] M. J. Latorre,[a] M. I. Fernández,[b] M. A. García,[c] S. García[a] & C. Herrero[a]*

[a]*Departamento de Química Analítica, Nutrición y Bromatología, Facultad de Ciencias de Lugo, Augas Férreas s/n, 27002 Lugo, Spain*
[b]*Departamento de Química Inorgánica, Facultad de Ciencias de Lugo, Augas Férreas s/n, 27002 Lugo, Spain*
[c]*Departamento de Toxicología y Legislación Sanitaria, Facultad de Ciencias de Lugo, Augas Férreas s/n, 27002 Lugo, Spain*

(Received 3 March 1995; accepted 24 March 1995)

Eleven legal parameters of quality of honey were determined in 29 honey samples divided into two categories: natural and processed honeys. Multivariate chemometric techniques such as principal component analysis, cluster analysis, linear discriminant analysis, KNN and SIMCA, are used to classify honeys on the basis of the chemical data. Using only two features, total acidity and diastase, a nearly correct classification was achieved.

INTRODUCTION

Classification of product brands and quality of origin is a very active area for the application of chemometric classification procedures (Forina & Lanteri, 1984; Brown *et al.*, 1992). The use of pattern recognition techniques in food chemistry problems has widely increased in recent years. After the works of Kwan and Kowalski on the classification of wines of *Vitis vinifera* cv. Pinot Noir from France and the United States (Kwan *et al.*, 1979; Kwan & Kowalski, 1980), a number of examples were reported in the literature concerning a variety of products in which several classes of chemical variables and sensory properties are considered and different statistical tools are used. Geographic classification of Italian olive oils (Derde *et al.*, 1984) and concentrated orange juices (Bayer *et al.*, 1980) were achieved using pattern recognition techniques such as cluster analysis, linear discriminant analysis, K-nearest neighbour and SIMCA. Principal component analysis, linear discriminant, quadratic discriminant and SIMCA methods were compared for classification and predictive ability in the separation of two species of fish (Franco *et al.*, 1990). Chemometric studies on minor and trace elements in cow's milk were realized to differentiate two types of milk according to the kinds of cow's feeding (Favretto *et al.*, 1994). Pattern recognition methods have been used in characterization and classification of wine and alcoholic beverages: Galician (Spain) wines by metal ions (Herrero *et al.*, 1994), Venetian (Italy) wines by inorganic and classical determinations plus aroma compounds (Moret *et al.*, 1994), Portuguese wines by free amino acid profiles (Vasconcelos & Chaves, 1989), French red wines by elements, amino acids and aromatic alcohols (Etiévant *et al.*, 1988) and German wines by volatile compounds, non-volatile acids and amino acids (Maarse *et al.*, 1987). Multivariate chemometric techniques were used to classify alcoholic distillates and to develop a typification model for Galician liquors, on the basis of data from chromatographic analyses (Cruz *et al.*, 1993).

In a previous work (Herrero & Peña, 1993), we showed that multivariate statistical methods can be used sucessfully to achieve a correct geographic classification of honey samples from different origin. In this work, an attempt was made to distinguish between natural and processed honeys. Processed honeys, due to their lower price, can be used as possible substrates for falsification of natural honeys. Heat treatment, generally realized to handle and to homogenize honeys from different origins in the bottle industry, produces alterations that can be detected by quality control parameters. Applications of pattern recognition methods to key components related to thermal changes in honey were used to achieve a differentiation between processed and industrially-commercialized honeys and natural honeys directly obtained from the producers.

MATERIALS AND METHODS

Honey samples

Sixteen natural honey samples were provided by the local association of beekepers with guaranteed origin and made by the traditional procedures in the producing region. All natural samples examined were unprocessed honeys of random (mixed) floral type of the 1992

*To whom correspondence should be addressed.

harvest. None of these samples underwent any process that could alter their composition. Thirteen processed honey samples were obtained from different supermarkets and commercial areas. Samples were collected in glass bottles and stored in darkness at 3–4°C until analysis.

Analytical determinations

Samples were analysed to determine the following: humidity, ash, insoluble matter, reducing sugars, sucrose, conductivity, free acidity, total acidity, pH, hydroxymethylfurfural (HMF) and diastase activity. All analyses were performed according to the official methods of Spanish legislation (A.O.A.C., 1990; BOE, 1986) and they have been described in detail in a previous work (Herrero & Peña, 1993).

Data analysis

Each honey sample (object) was considered as an assembly of twelve variables represented by the chemical data. These variables called 'features' formed a 'data vector' which represented a honey sample. Data vectors belonging to the same group, such as processed or natural, were analysed. The group was then termed a 'category'. Pattern recognition tools used in this work were as follows.

Autoscale
This is the most widely used scaling technique (Kowalski & Bender, 1972). The procedure standardizes a variable m according to:

$$Y_{mj} = \frac{(x_{mj} - \bar{x}_m)}{s_m}$$

where Y_{mj} is the value j for the variable m after scaling, x_{mj} is the value j for the variable m before scaling, $\bar{x}_m$ is the mean of the variable and s_m is the standard deviation of the variable. The result is a variable with zero mean and a unit standard deviation.

Feature selection
Selection of variables containing the most powerful information to achieve a correct classification between honey samples of the two considered categories was made by generating a set of decorrelated variables on the basis of their variance weights (Meloun *et al.*, 1992*a*).

Cluster analysis
Clustering techniques is an unsupervised classification procedure that involves a measurement of either the distance or similarity between objects to be clustered. Objects are grouped in clusters in terms of their nearness or similarity. The initial assumption is that the nearness of objects in the p-space defined by the variables reflects the similarity of their properties (Massart & Kaufman, 1983).

PCA (Principal component analysis)
This procedure (Mardia *et al.*, 1979) was used mainly to achieve a reduction of dimensionality, i.e. to fit a j-dimensional subspace to the original p-variate ($p > > j$) space of the objects, and it allows a primary evaluation of the between-category similarity.

LDA (Linear discriminant analysis)
This classification procedure (Wold *et al.*, 1984) maximizes the variance between categories and minimizes the variance within categories. The method renders a number of orthogonal linear discriminant functions, equal to number of categories minus 1.

KNN (K nearest neighbour)
This classification method, based on the distance between objects in the p-space as its criterion (Wold *et al.*, 1984), is used to classify an object in the category which contributes with the greatest number of the K nearest known objects. It is a non-parametric method inasmuch as it does not formulate a hypothesis on the distribution of the variables used. Only the closest K objects are used in making any given classification. The importance of a given feature in making the decisions is proportional to its contribution to the distance calculation. The inverse square of Euclidean distance was used in this work.

SIMCA (soft independent modelling of class analogy)
This classification procedure uses linear discriminant functions derived from disjointed principal component analysis of the data (Wold, 1976). One set of functions is derived for each category studied by computing the category-mean and a specified number of the principal components. Objects are classified into the category whose principal component model best reproduces the data. Only data points which are members of a given category are used in determining the model functions for that category. The importance of each feature in classification is determined by its contribution to the category covariance matrices.

The data analysis was performed in few steps.

(1) Preliminary data analysis by cluster and principal component analysis using the complete data set.

(2) Classification techniques LDA, KNN and SIMCA were applied to the complete data set with all samples included in the training set.

(3) For practical reasons it is important to know the minimum number of features needed to obtain a correct classification. This could be achieved by choosing features which contained the most discriminant information for the classification. The criterion used for selection was the variance weights.

(4) The reliability of the classification previously obtained was checked. The 29 objects were randomly divided between training (or learning) set and evaluation (or prediction) set. LDA, KNN and SIMCA were applied, based only on the features selected in Step (3).

Pattern recognition analyses were performed by means of statistical software packages StatGraphics

Table 1. Results of quality control parameters determined

	Processed honey samples ($n = 13$)			
	Mean	SD	Max.	Min.
Processed honey samples ($n = 13$)				
Reducing sugars (%)	76.4	5.75	93.4	70.0
Insoluble matter(%)	0.287	0.309	0.820	0.00
Sucrose (%)	3.53	3.87	14.79	0.55
Conductivity (μS cm^{-1})	250.3	80.9	367.0	129.0
Free acidity (meq kg^{-1})	21.4	5.89	36.4	13.9
Total acidity (meq kg^{-1})	21.45	5.93	36.41	13.86
pH	3.99	0.22	4.45	3.61
HMF(mg kg^{-1})	41.9	56.1	219	12.9
Ash (%)	0.13	0.09	0.38	0.06
Humidity (%)	16.75	1.86	19.80	13.60
Diastase (u Gothe)	13.0	8.14	31.8	3.58
Natural honey samples ($n = 16$)				
Reducing sugars (%)	71.7	1.56	74.0	68.3
Insoluble matter(%)	0.018	0.021	0.080	0.000
Sucrose (%)	1.12	1.06	4.26	0.33
Conductivity (μS cm^{-1})	302.9	39.8	401.0	231.0
Free acidity (meq kg^{-1})	35.9	6.20	47.0	28.9
Total acidity (meq kg^{-1})	39.9	6.54	49.9	30.6
pH	4.02	0.34	4.64	3.38
HMF(mg kg^{-1})	8.74	3.22	16.3	2.88
Ash (%)	0.26	0.11	0.45	0.05
Humidity (%)	17.74	0.92	19.70	16.60
Diastase (u Gothe)	27.3	10.4	48.1	14.1

SD, standard deviation; Max, maximum; min., minimum.

(Statistical Graphics Corporation, 1991) and Parvus (Forina *et al.*, 1988) in a Gulf-Tech 486/50 computer using a HP Laserjet 4ML as graphic output.

RESULTS AND DISCUSSION

The results of the 11 quality control parameters determined are summarized in Table 1. The levels obtained in natural honey were similar to those found by other authors (Huidobro, 1983; Sancho, 1990; Herrero & Peña, 1993) in honey samples from Galicia (NW Spain). Processed honey samples present levels of acidity, HMF and diastase higher that natural samples.

The search for natural groupings in the samples is one of the main ways to study the structure of the data. The cluster analysis describes the nearness between honey samples using, in this case, the squared Euclidean distance between objects. It can as well represent the multidimensional space by mapping it on two dimensions. Thus, a similarity matrix S_{29x29} was constructed from the autoscaled data, the elements of this similarity matrix were the squared Euclidean distance of one object from the rest. The clustering procedure used was the Ward method. This aglomerative method considers in each step the heterogeneity or deviance (sum of the squares of the distance of an object from the barycentre of the cluster) of every possible cluster that can be created by linking two existing clusters (Meloun *et al.*, 1992*b*). The results obtained showed the presence of honey clusters; the data of quality control parameters contained significant information to achieve a two-category classification between natural and processed honeys. The results of cluster analysis are shown as a dendrogram in Fig. 1. At a similarity level of 0.5, five clusters were found, which can be identified as follows. The first cluster was composed of 10 processed honeys; the second and third clusters contained only one processed honey; the fourth cluster was formed by eight

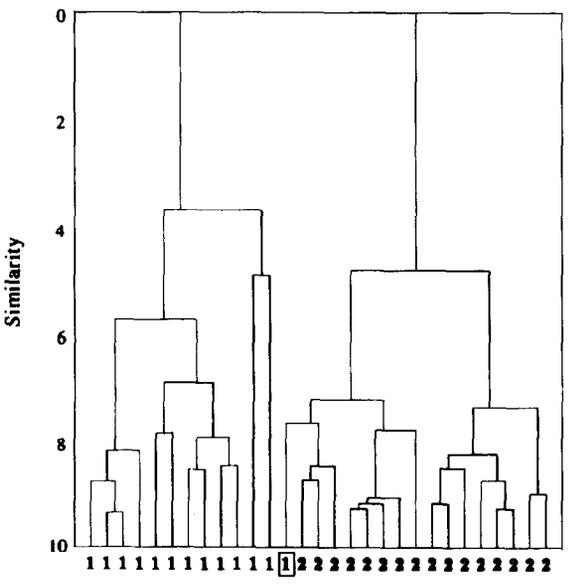

Fig. 1. Dendrogram of cluster analysis. 1, processed honey; 2, natural honey.

natural plus one processed honeys; the last cluster was made up of eight natural honeys. Examination of the dendrogram at a similarity level of 0.7 provided similar results to a 0.5 similarity level. One processed honey sample was grouped in a cluster with a high similarity level of natural honeys.

Principal components analyses were performed using Statgraphics. Each principal component or eigenvector is orthogonal and it is a linear combination of the original variables. From the loadings of features in the first and second eigenvectors, diastase and total acidity are the dominating features in the first principal component that represent 37.47% of total variability, while insoluble matter dominates in the second principal component that represents 17.18% of total variability. Chemically the first component can be associated with the acidity of the honey and the second principal component with its solids content. The first three eigenvectors account for the 67.28% of the total variability. Examining a three-dimensional plot of the samples in the space defined by the three first principal components (see Fig. 2), a natural separation between processed and natural honeys was found. In this factor space, natural honeys formed a group that included one sample of processed honey. This result is consistent with the conclusions obtained by cluster analysis, where one sample of processed honey appeared in a cluster formed by natural honeys.

The three classification methods considered above, LDA, KNN and SIMCA, were applied to an initial matrix containing the 29 objects and 11 variables divided between natural and processed honeys [Step (2)]. In this case, all samples were in the training set. The original data were autoscaled to eliminate the effect of the diverse size of the variables. The recognition ability for the two classes was highly satisfactory, all natural samples were correctly classified, two samples of processed honeys were misclassified. Similar percentage of good classification were obtained by KNN in the complete data set using inverse square of the Euclidean distance and K = 3. The same results were achieved when KNN with K = 4 and K = 5 were applied (see Table 2). Using SIMCA 82.7% of the honey samples were assigned to their group (five natural honeys were misclassified).

In Step (3), a selection of the minimum number of variables to reach a correct classification was performed. This could only be achieved by choosing the features which contained the most discriminant information for the classification. The selection of a small number of key variables increases the reliability of mathematical classification, eliminates features with minor information and allows a visual examination of the data set by two-dimensional plot of the key features. The criterion used for the variable selection was the variance weight that is a quantitative estimate of the utility of a given measurement for separating categories (Meloun *et al.*, 1992*a*). In the case of two categories, with a data matrix X_{IM} portioned in two matrices, one for each category:

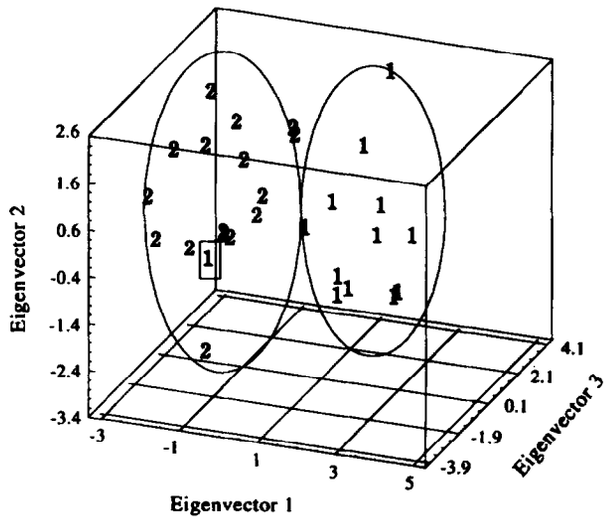

Fig. 2. Eigenvector projection of honey samples. 1, processed honey; 2, natural honey.

$$X_{IM} = \begin{bmatrix} X_{I_1 M} \\ X_{I_2 M} \end{bmatrix}$$

the estimate of the variance of the variable m within the category g is:

$$V_{gm} = (I_g - 1)^{-1} \sum_{g=1}^{I_g} (x_{i_g m} - \bar{x}_{gm})^2$$

and the intercentroid variance between category 1 and category 2 is estimated by:

$$V_{cm} = (\bar{x}_{1m} - \bar{x}_m)^2 + (\bar{x}_{2m} - \bar{x}_m)^2$$

The Fisher weight FW_m of the variable is:

$$FW_m = 4 \frac{V_{cm}}{(V_{1m} + V_{2m})}$$

In the case of more than two categories, FW becomes the average of the weights computed for each possible pair of categories. The variance weight (VW) is also commonly used, and in the case of two categories, VM_m is given by:

$$VM_m = \frac{(1 + FW_m)}{2}$$

Table 2. Classification with LDA, KNN and SIMCA using all features

Method	Percentage correct assignation (processed/natural)	Misclassified samples
LDA	84.6/100	2
KNN		
K = 3	84.6/100	2
K = 4	84.6/100	2
K = 5	84.6/100	2
SIMCA	100/68.8	5

For more than two categories VW is the geometrical average of the pairwise variance weights. The easiest way of feature selection is to evaluate separately on its discriminating ability expressed by the variance weight. In this way, the variables can be arranged according to their importance to separate categories. Total acidity, free acidity and diastase were found to be the features containing most powerful and discriminant information in order to achieve a separation and classificaction between processed and natural honeys. However, when more than one variable has to be selected this method was not optimal. If two correlated variables are considered, both will appear with a similar VW and the variable selected in the second place does not contribute with different discriminating information to the variable selected in the first place. In this case, total and free acidity are strongly correlated. Therefore, if total acidity is selected first, little aditional information is provided by selecting free acidity in the next step. For these reasons, a more sophisticated method that generates a set of decorrelated (orthogonal) variables was used. The new variables are obtained through successive decorrelations of original variables in the order corresponding to their variance weights. If the first variable to be decorrelated is denoted with m, then the other variables are modified accordingly:

$$x_v^* = A x_v + B x_m$$

where

$$A = \frac{1}{\sqrt{(1 - r_{vm})}}$$

and

$$B = -\frac{S_v^2 r_{vm}}{S_m^2 \sqrt{(1 - r_{vm})}}$$

r_{vm} is the correlation coefficient between variables v and m, the variables x_v^* obtained is uncorrelated with m and its variance is s_v^2 before correlation. The variance weights of the variables x_v^* are then computed, and these new variables are used to determine the next variable decorrelated (Forina *et al.*, 1988). The first variable selected was total acidity, the second diastase and third insoluble matter. These results agree with those obtained by PCA, where total acidity and diastase were the dominant features in the eigenvector 1 and insoluble matter in eigenvector 2.

Finally, in Step (4), the reliability of the classification using only the two selected features, total acidity and diastase, was tested. The honey samples were randomly divided between training (or learning) set and evaluation (or prediction) set. The percentage of objects placed in the evaluation set was 25%. Such division allows us to have a sufficient number of samples in the training set as well as a representative number of honeys in the evaluation set. In order to reach a good test of the recognition and prediction ability of each method and to see how well the data could be classified quatitatively, the previous division procedure between training and evaluation sets was repeated 10 times to obtain 10 files with different contitutions of the two sets. LDA, KNN and SIMCA were applied to the 10 considered files. The number of neigbours employed in KNN was K = 3, because previously [Step (2)], the same percentage of good classification was obtained as that using K = 4 or 5. The results obtained are presented in Table 3. Reducing the number of features to two, LDA and KNN, showed essentially the same results. A high level of correct assignation of natural honeys with a percentage of successes in recognition and prediction between 95 and 100% was achieved. For processed honeys the percentage of correct classification was less successful. This fact indicates that the pattern recognition procedures are selective for the natural honeys; the probability of a natural honey being classified as processed is virtually nil. However, a minor level of hits in classification and prediction of processed honeys suggests that there exists a certain probability that a processed honey might be classified as natural. These results coincide with the ones obtained by PCA and cluster analysis, where one processed honey sample was grouped into natural samples. Results achieved when SIMCA was applied were less successful; recognition ability was about 88% and prediction ability was 73.3% for processed honeys and 90.0% for natural honeys. In this case, also, a larger percentage of correct asignation and prediction was reached for the natural honey samples.

The data were then examined by a two-dimensional plot of two key features. A plot of total acidity vs diastase is shown in Fig. 3; one sample of processed honey was grouped as natural confirming the possibility noted earlier.

CONCLUSION

Pattern recognition is able to extract useful information for an amount of data. Information was used to relate chemical composition of honeys with their processing. Differentiation and classification of processed and unprocessed honey samples was made possible by using two quality control parameters: total acidity and diastase and applying multidimensional chemometric

Table 3. Classification with LDA, KNN and SIMCA using only two selected features: total acidity and diastase

Category	Percentage recognition ability	Percentage prediction ability
LDA		
Processed honey	93.5	86.9
Natural honey	100.0	100.0
KNN (K = 3)		
Processed honey	94.0	86.7
Natural honey	95.1	100.0
SIMCA		
Processed honey	88.0	73.3
Natural honey	88.3	90.0

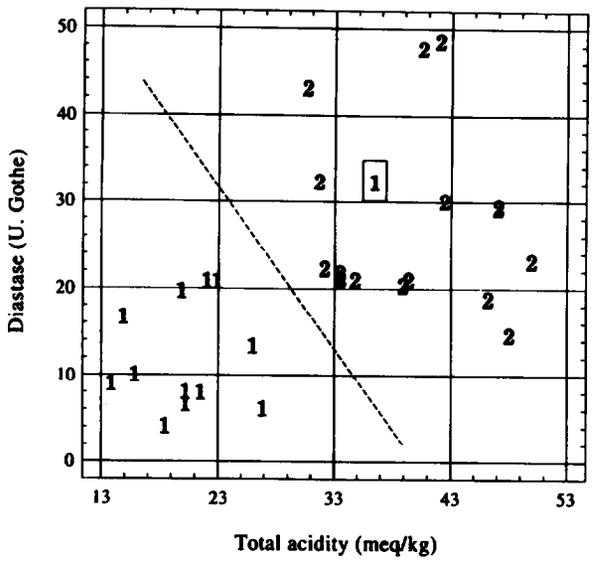

Fig. 3. Plot of the two first-selected features, total acidity vs second selected feature, diastase. 1, processed honey; 2, natural honeys.

techniques. A very low diastase activity and an appreciable amount of HMF have been considered as indicative that honey has been subjected to unfavourable temperature conditions. However, in this work we can demonstrate that acidity was also an important feature to achieve a correct classification and that diastase and HMF are redundant information for this purpose. Use of all available features is unnecessary and undesirable, as the use of variables with no extra discriminating information only introduces noise in the pattern recognition process. The selected features can be used to obtain a highly successful classification. But, if a processed honey sample undergoes a controlled thermal process that does not alter certain quality control values, this honey can be classified as natural.

ACKNOWLEDGEMENTS

We express our gratitude to the Asociación Lucense de Apicultura for providing natural honey samples. We are grateful to A. Chouciño for valuable observations. This work was partly financed by the Instituto Lucense de Desarrollo, Lugo, Spain.

REFERENCES

A.O.A.C. (1990). *Official Methods of Analysis*, ed. K. Helrich. Association of Analytical Chemists, Arlington.

Bayer, S., McHard, J. A. & Winefordner, J. (1980). Determination of geographic origin of frozen concentrated orange juices via pattern recognition. *J. Agric. Food Chem.*, **28**, 1306.

BOE (1986) 'Orden de 12 de Junio de 1986 por la que se aprueban los métodos oficiales de análisis para la miel'. Boletín Oficial del Estado, Madrid, 18 Junio.

Brown, S. D., Bear, R. S. & Blank, T. B. (1992). Chemometrics. *Analyt. Chem.*, **64**, 22R–49R.

Derde, M. P., Coomans, D. & Massart, D. L. SIMCA (1984). Demonstrated with characterization of italian olive oils. *J. Assoc. Off. Analyt. Chem. Symp. Ser.*, **18**, 49–52.

Cruz, M., Sáez, J. A. & López-Palacios, J. (1993). Typification of alcoholic distillates by multivariate techniques using data from chromatographic analyses. *Analyst*, **118**, 801–5.

Etiévant, P., Schlich, P., Symonds, P., Bouvier, J. C. & Bertrand, A. (1988). Varietal and geographic classification of french red wines in terms of elements, amino acids and aromatic alcohols. *J. Sci. Food Agric.*, **45**, 25–41.

Favretto, L., Vojnovic, D. & Campisi, B. (1994). Chemometric studies on minor and trace elements in cow's milk. *Analyt. Chim. Acta*, **293**, 295–300.

Forina, M. & Lanteri, S. (1984). Data analysis in food chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*, ed. B. R. Kowalski. Riedel Publishing, Dordrecht, pp. 305–51.

Forina, M., Leardi, R., Armanino, C. & Lanteri, S. (1988). *Parvus: An Extendable Package of Programs for Data Exploration, Classification and Correlation.* Elsevier, Amsterdam.

Franco, M. A., Seeber, R., Sferlazzo, G. & Leardi, R. (1990). Classification and prediction ability of pattern recognition methods applied to sea-water fish. *Analyt. Chim. Acta*, **233**, 143–7.

Herrero, C. & Peña, R. (1993). Pattern recognition analysis applied to classification of honeys from two geographics origins. *J. Agric. Food Chem.*, **41**, 560–4.

Herrero, C., Latorre, M. J., García, C. & Mèdina, B. (1994). Pattern recognition analysis applied to classification of wines from Galicia (NW Spain) with certified brand of origin. *J. Agric. Food Chem.*, **42**, 1451–5.

Huidobro J. F. (1983). La miel: algunos parámetros de interés en su control de calidad. Ph.D. thesis, Universidad de Santiago de Compostela, Spain.

Kowalski, B. R. & Bender, C. F. (1972). Pattern recognition. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc.*, **94**, 5632–9.

Kwan, W. O., Kowalski, B. R. & Skogerboe, R. K. (1979). Pattern recognition analysis of elemental data. Wines of *Vitis vinifera* cv. Pinot Noir from France and the Unites States. *J. Agric. Food Chem.*, **27**, 1321–6.

Kwan, W. O. & Kowalski, B. R. (1980). Pattern recognition analysis of gas chromatographic data. Geographic classification of wines of *Vitis vinifera* cv. Pinot Noir from France and the Unites States. *J. Agric. Food Chem.*, **28**, 356–9.

Maarse, H., Slump, P., Tas, A. C. & Schaefer, J. (1987). Classification of wines according to type and region based on their composition. *Zt. Lebensm. Unters Forsch.*, **184**, 198–203.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, New York.

Massart, D. L. & Kaufman, L. (1983). Hierarchical clustering methods. In *The Interpretation of Analytical Data by Use of Cluster Analysis.* Wiley, New York.

Meloun, M., Militky, J. & Forina, M. (1992a). Scaling, weighting, transforms. In *Chemometrics for Analytical Chemistry.* Ellis Horwood, New York, pp. 214–26.

Meloun, M., Militky, J. & Forina, M. (1992*b*). Clustering. In *Chemometrics for Analytical Chemistry.* Ellis Horwood, New York, pp. 244–69.

Moret, I., Scarponi, G. & Cescon, P. (1994). Chemometric characterization and classification of five Venetian white wines. *J. Agric. Food Chem.*, **42**, 1143–53.

Statistical Graphics Corporation. (1991). *Statgraphics: Reference Manual.* STSC Inc., Rockville.

Sancho, M. T. (1990). Estudio de las mieles producidas en la Comunidad Autónoma del País Vasco. Ph.D. thesis, Universidad de Santiago de Compostela, Spain.

Vasconcelos, P. & Chaves, H. (1989). Classification of elementary wines of Vitris Vinifera varieties by pattern recognition of free amino acid profiles. *J. Agric. Food Chem.*, **37**, 931–7.

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. *Pattern Recognit.*, **8**, 127–39.

Wold, S., Albano, C., Dun, W. J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johanson, E., Lindberg, W. & Sjöstrom, M. (1984). Multivariate data analysis in chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*, ed. B. R. Kowalski. Riedel Publishing, Dordrecht, pp. 17–97.