Food Chemistry 66 (1999) 263–268

Food Chemistry

www.elsevier.com/locate/foodchem

Analytical, Nutritional and Clinical Methods Section

Chemometric classification of honeys according to their type. II. Metal content data

M.J. Latorre, R. Peña, C. Pita, A. Botana, S. García, C. Herrero *

Departamento de Química Analítica, Nutrición y Bromatología, Facultad de Ciencias de Lugo, Augas Férreas s/n, Campus Universitario, 27002 Lugo, Spain

Received 21 April 1998; received in revised form and accepted 22 September 1998

Abstract

The objective of this work was to develop a method to confirm the geographical authenticity of Galician-labelled honeys as Galician-produced honeys. Eleven metals were determined in 42 honey samples divided into two categories: natural Galician honeys and processed non-Galician honeys. Multivariate chemometric techniques such as principal component analysis, linear discriminant analysis, KNN and SIMCA are used to classify honeys according to their type and origin on the basis of the chemical data. Using only three features, Cu, Mn and Li, an almost correct classification was achieved. © 1999 Elsevier Science Ltd. All rights reserved.

1. Introduction

In recent years, pattern recognition techniques have been widely applied in food chemistry (Brown, Bear, & Blank, 1992; Forina & Lanteri, 1984). After the work of Kwan et al. on the classification of wines of *Vitis vinifera* cv. Pinot Noir from France and the United States (Kwan & Kowalski, 1980, Kwan, Kowalski, & Skogerboe, 1979), a number of examples were reported in the literature concerning a variety of products. These include the geographic classification of olive oils (Derde, Coomans, & Massart, 1984), concentrate orange juices (Bayer, McHard, & Winefordner, 1980) and wines (Herrero & Médina, 1990). More recently, there has been varied work based on diverse types of chemical variables and sensory properties using different statistical tools: 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, Seeber, Sferlazzo, & Leardi, 1990); chemometric studies on minor and trace elements in cow's milk were made to differentiate two types of milk according to feeding (Favretto, Vojnovic, & Campisi, 1994); pattern recognition methods have been used in characterization and classification of wine and alcoholic beverages (Etiévant, Schlich, Symonds, Bouvier, & Bertrand, 1988; Herrero, Latorre, García, & Médina, 1994; Maarse, Slump, Tas, & Schaefer, 1987; Moret, Scarponi, & Cescon, 1994; Vasconcelos & Chaves, 1989).

In a previous work (Herrero & Peña, 1993), we showed that multivariate statistical methods applied to physicochemical parameters can be sucessfully used in order to achieve a correct geographic classification of honey samples from different origins. Also, pattern recognition procedures were used to obtain a classification model based on quality control data to distinguish between natural and processed honey samples (Herrero et al., 1996). In none of these cases were pollen studies necessary to obtain an almost correct assignation of the honey samples.

In this work, we present the results of the application of pattern recognition methods to key metals data in honey to differentiate between processed and industrially commercialized honeys from various origins and natural honeys from Galicia directly obtained from the producers. The basis of the classification procedure was the metal composition of honey samples. Minerals seem to be good candidates for a classification system, as they are stable; however, the number of research papers in the literature concerning the metal content of honey is small (Rodríguez-Otero, Paseiro, Simal, Terradillos, & Cepeda, 1992; Stein & Umland, 1986; Li, Whadat, & Neeb, 1995) and only in one case was the data obtained

* Corresponding author. E-mail: cherrero@lugo.usc.es

0308-8146/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0308-8146(98)00217-9

used to carry out a chemometric classification of honey samples according to their geographical origin (Feller, Vincent, & Beaulieu, 1989). The interest in this classification model is two-fold. On the one hand, processed honeys undergo a heat treatment, generally targeted to handle and homogenize honeys from different origins in the bottle industry; this process causes alterations that can affect the properties and quality of the product. On the other hand, processed honeys from various origins, due to their lower price, can be used as possible substrates for falsification of natural Galician honeys.

2. Materials and methods

2.1. Honey samples

Twenty-two natural honey samples from Galicia were provided by the local association of beekepers with guaranteed origin and made by traditional procedures in the producing region. All the natural samples examined were unprocessed honeys of random (mixed) floral type. None of these samples underwent any process that could alter their composition. Twenty processed honey samples from various origins except Galicia were obtained from different supermarkets and commercial areas. Samples were collected in glass bottles and stored in the dark at 3–4°C until analysis.

2.2. Analytical determinations

Eleven selected metals, Li, Rb, Na, K, Mg, Zn, Cu, Fe, Mn, Ni, and Co were measured using a Varian AA 10 Plus atomic spectrometer. Li, Rb, Na, and K were determined by atomic emission spectroscopy and, the remaining elements were determined by atomic absorption spectroscopy. Analytical procedures have been described in detail in previous works (Herrero & Peña, 1993; Rodríguez-Otero et al., 1992). All determinations were made twice.

2.3. Data analysis

Each honey sample (object) was considered as an assembly of 11 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 (processed or natural) were analyzed. The group was then termed a "category". Pattern recognition tools used in this work were as follows.

2.3.1. 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.

2.3.2. Feature selection

Selection of variables containing the most powerful information for a correct classification of honey samples of the two considered categories was carried out on the basis of Stepwise Bayesian Analysis (Forina, Leardi, Armanino, & Lanteri, 1988).

2.3.3. Principal component analysis (PCA)

This procedure (Mardia, Kent, & Bibby, 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 pit allows a primary evaluation of the between-category similarity.

2.3.4. Linear discriminant analysis (LDA)

This classification procedure (Wold et al., 1984) maximizes the variance between categories and minimizes the variance within categories. The method operates in a p-space by calculating a p-1 dimensional surface which separates the two categories as well as possible.

2.3.5. K nearest neighbour (KNN)

This 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 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 to make any given classification. The importance of a given feature in taking the decisions is proportional to its contribution to the distance calculation. The inverse square of Euclidean distance was used in this work.

2.3.6. Soft independent modelling of class analogy (SIMCA)

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 the 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.
4. The reliability of the classification obtained before was checked. The 42 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 analysis were performed by means of statistical software packages STATGRAPHICS (Statgraphics, 1991) and PARVUS (Forina et al., 1988) in a Gulf-Tech Pentium/150 MHz computer using a HP Laserjet 4ML as graphic output.

3. Results and discussion

The results of the 11 metals determined in honey samples are summarized in Table 1.

The levels obtained in natural Galician honey were similar to those found by other authors in honey samples from Spain (Huidobro, 1983; Sancho, 1990; Serra, 1989) and also in honey samples from Galicia (Rodríguez-Otero et al., 1992). The contents of Co and Ni were, in all cases, less than detection limits, 0.02 and 0.05 μg g^{-1}, respectively. The variability of results for all metals in both groups are large, reflecting the wide variation in the composition of honey samples. A higher content of minerals was found in Galician honeys. This might be due to the high acidity of Galician honeys (Herrero et al., 1996). The levels of Mn have special interest, the great difference in concentration for this metal in natural honeys from Galicia (5.2 μg g^{-1}) and processed honeys from various origins (1.1 μg g^{-1}) can be an indicator for geographical classification of samples. This result is consistent with the conclusions of a previous work (Herrero & Médina, 1990) where we demonstrated that the manganese content also plays a key role in geographical differentiation between Galician wines and the wines from other Spanish areas.

Table 1
Results of metals determined. All results are in ppm except Li in ppb

	Natural Galician honeys ($n=22$)		Industrial non-Galician honeys ($n=20$)	
	Mean	SD	Mean	SD
Li	9.2	6.4	3.2	1.1
Na	115	52.8	54.2	16.2
K	1345	665	618	522
Rb	1.5	1.2	0.3	0.7
Mg	77.0	43.4	105	51.0
Zn	2.0	1.3	1.5	0.7
Fe	3.7	1.7	3.8	3.0
Mn	5.2	3.1	1.1	1.1
Cu	0.89	0.73	0.14	0.10
Ni	<0.02	–	<0.02	–
Co	<0.05	–	<0.05	–

The search for natural groupings in the samples is one of the main ways to study the structure of the data. Principal components analysis (PCA) was used to provide a partial visualization of data in a reduced-dimension plot. PCA were performed using Statgraphics. The principal components or eigenvectors are orthogonal and they are a linear combination of the original variables. From the loadings of features in the first and second eigenvectors (Table 2), lithium, potassium, rubidium, manganese and copper are the dominant variables in the first principal component that represent 41.36% of the total variability; while iron, sodium and zinc dominate in the second principal component that represents 22.07% of total variability. The first three eigenvectors account for the 75.52% of the total variability. Examining a three dimensional plot of the samples in the space defined by the three first principal components (Fig. 1), a natural separation between natural honeys from Galicia and processed honeys from non-Galician origin was found. In this factor space, natural honeys formed a group that includes two samples of processed honey.

The three classification methods considered above, LDA, KNN and SIMCA, were applied to an initial matrix containing the 42 objects and 9 variables divided between natural Galician and processed non-Galician honeys (step 2). All variables are previously transformed according to the autoscale procedure described above in order to avoid the effect of their different size and to put all of them on equal footing in terms of their variance (Sharaf, Illman, & Kowalski, 1986). 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. Using LDA the recognition ability for the two classes was highly satisfactory; all samples were correctly classified (Table 3).

Also a good percentage of correct classification was obtained by KNN in the complete data set using the inverse square of the Euclidean distance and $K=2$. The level of classification achieved was 95.4% for the natural Galician group and 95.0% for the processed

Table 2
Loadings of the variables in the first two principal components

	Variable								
	Li	Na	K	Rb	Mg	Zn	Fe	Mn	Cu
Principal component 1	0.396	0.301	0.439	0.430	0.165	0.224	0.134	0.422	0.409
Principal component 2	−0.199	−0.469	0.213	0.144	0.395	−0.534	−0.421	0.182	−0.145

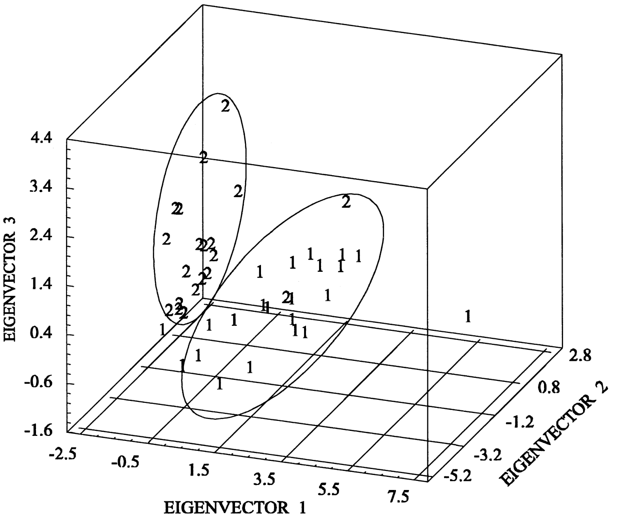

Fig. 1. Eigenvector projection of honey samples. 1: natural Galician honey, 2: processed non-Galician honey.

Table 3
Classification with LDA, KNN and SIMCA using all features

Method	% correct assignation (natural Galician/ processed non-Galician)	Misclassified samples
LDA	100/100	0
KNN		
K=2	95.4/95.0	2
K=3	90.9/90.0	4
K=4	90.9/90.0	4
K=5	90.9/90.0	4
SIMCA	100/90.0	2

non-Galician group (one sample was misclassified in both classes). When the number of neighbour employed was *K*= 3, a less sucessful result was obtained: The level of correct classification achieved in two groups considered was 90.9 and 90.0%, respectively. The same results were achieved when KNN with *K*= 4 and *K*= 5 were applied. Using SIMCA, 95.2% of the honey samples were assigned to their group (two processed non-Galician 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 a two-dimensional plot of the key features. The method used for the variable selection was stepwise bayesian analysis (SBA) (Forina et al., 1988). This method carries out a feature selection procedure based on Bayesian analysis; classification is performed by assuming an underlying multivariate normal distribution for every variable i in every category (or group) g. The probability density $p(x/g)$ for objects in every category, and with the a-posteriori probability $p(g/x)$, corrected by a loss factor, are computed. The efficiency of the variables in the classification of the objects into class g is the mean of $p(g/x)$ computed on every object in the class. The efficiency can be estimated from the classification rate (percentage of misclassified samples) or directly computed. In this case, when the SBA method was applied Li, Cu and Mn were selected as the variables that produce the best classification of the samples between the two considered groups, natural Galician and processed non-Galician honeys. The percentage of misclassified samples were as follows: Li 2.38%, Cu 9.52% and Mn 16.67%. The remaining variables produced a less successful efficiency in the classification; as a consequence of the variable selection procedures applied, Li, Cu and Mn were the features selected for further chemometric techniques.

Finally, in step 4, the reliability of the classification using only the three selected features, Li, Mn and Cu, 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 by random selection in each category was 20% (four samples of natural Galician origin and four samples of processed non-Galician origin). 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 satisfactory test of the recognition and prediction ability of each method and to see how well the data could be classified quantitatively, the previous division procedure between training and evaluation sets was repeated five times to obtain five files with different constitutions of the two sets. LDA, KNN and SIMCA were applied to the five mentioned files. The number of neighbours

Table 4
Classification with LDA, KNN and SIMCA using only three selected features: Li, Cu and Mn

Category	% recognition ability	% prediction ability
LDA		
Natural Galician honey	84.6	84.2
Processed non-Galician honey	95.0	90.0
KNN ($K=3$)		
Natural Galician honey	83.3	90.0
Processed non-Galician honey	90.2	95.0
SIMCA		
Natural Galician honey	100	100
Processed non-Galician honey	76.2	80.1

employed in KNN was $K=3$, because previously (step 2), the same percentage of correct classification was obtained when using $K=4$ or 5. In this case, KNN with $K=2$ was not used because it probably produces a higher level of correct assignation but a poor validation of the structure of data due to a reduced number of neighbours considered. The results obtained are presented in Table 4. Reducing the number of features to three, LDA and KNN showed essentially the same good results. A high level of correct assignation of natural honeys from Galicia with a percentage of success in recognition and prediction between 83.3 and 100% was achieved. For processed honeys, the percentage of correct classification was also successful when LDA and KNN were applied. A less successful result was achieved when SIMCA was employed for the classification of processed non-Galician honey samples. This fact indicates that this pattern recognition procedure is more selective for the natural honeys; the probability of a natural honey being classified as processed is low. 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, where two processed non-Galician honey samples were grouped into natural Galician samples.

4. Conclusion

We have demonstrated that 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 and geographic origins. Differentiation and classification of processed and unprocessed honey samples from Galician and non-Galician origin was made possible by using the concentration data of various selected metals and applying multidimensional chemometric techniques. The use of all available features is unnecessary and undesirable, because the consideration of variables with no extra discriminating information only introduces noise in the pattern recognition process. Employing the three selected features (Li, Mn and Cu) a highly successful classification between the two honey groups considered was achieved; pollen studies to obtain geographical classification were avoided and thus considerable time and effort were saved.

Because of their high quality, natural honeys from Galicia reach a high price in the market; therefore processed honey from a non-Galician origin can be used as substrates for adulteration or falsification due to their lower price and similar organoleptic properties. We can demonstrate that the metal content of honey combined with chemometric techniques can be used as a possible way for detection of frauds.

Acknowledgements

We express our gratitude to the Asociación Lucense de Apicultura for providing natural honey samples. This work was partly financed by the Instituto Lucense de Desarrollo (INLUDES), Lugo, Spain.

References

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

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

Derde, M. P., Coomans, D., & Massart, D. L. (1984). SIMCA demonstrated with Characterizacion of Italian olive oils. *Journal of the Association of Official Analytical Chemists Symposium Series, 18*, 49–52.

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. *Journal of the Science of Food and Agriculture, 45*, 25–41.

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

Feller, M., Vincent, B., & Beaulieu, F. (1989). Teneur en minraux et origine géographique des miels du Canada. *Apidology, 20*, 77–91.

Forina, M., & Lanteri, S. (1984). Data analysis in food chemistry. In B. R. Kowalski (Ed.), *Chemometrics, mathematics and statistics in chemistry* (pp. 305–351). Dordrecht, The Netherlands: Riedel Publishing.

Forina, M., Leardi, R., Armanino, C., & Lanteri, S. (1988). *Parvus: an extendable package of programs for data exploration, classification and correlation*. Amsterdam: Elsevier.

Franco, M. A., Seeber, R., Sferlazzo, G., & Leardi, R. (1990). Classification and prediction ability of pattern recognition methods applied to sea-water fish. *Analytica Chimica Acta, 233*, 143–147.

Herrero, C., Latorre, M. J., García, C., & Médina, B. (1994). Pattern recognition analysis applied to classification of wines from Galicia (NW Spain) with certified brand of origin. *Journal of Agricultural and Food Chemistry, 42*, 1451–1455.

Herrero, C., López, B., Latorre, M., García, M., García, S., & Fernández, M. (1996). Chemometric classification of honeys according to their type based on quality control data. *Food Chemistry, 55*, 281–287.

Herrero, C., & Médina, B. (1990). Utilisation de quelques éléments minéraux dans la differenciation des vins de Galice de ceux d'autres régions d Espagne. *Journal International des Sciences de la Vigne et du Vin, 24*, 147–156.

Herrero, C., & Peña, R. (1993). Pattern recognition analysis applied to classification of honeys from two geographics origins. *Journal of Agricultural and Food Chemistry, 41*, 560–564.

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. *Journal of the American Chemical Society, 94*, 5632–5639.

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 United States. *Journal of Agricultural and Food Chemistry*, 28, 356–359.

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 United States. *Journal of Agricultural and Food Chemistry, 27*, 1321–1326.

Li, Y., Whadat, F., & Neeb, R. (1995). Digestion-free determination of heavy metals (Pb, Cd and Cu) in homey samples using anodic-stripping differential pulse voltammetry and potentiometric-stripping analysis. *Fresenius Journal of Analytical Chemistry, 351*, 678–682.

Maarse, H., Slump, P., Tas, A. C., & Schaefer, J. (1987). Classification of wines according to type and region based on their composition. *Zeitschrift fucaer Lebensmitteluntersuchung und - forschung, 184*, 198–203.

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

Moret, I., Scarponi, G., & Cescon, P. (1994). Chemometric characterization and classification of five Venetian white wines. *Journal of Agricultural and Food Chemistry, 42*, 1143–1153.

Rodríguez-Otero, J., Paseiro, P., Simal, J., Terradillos, L., & Cepada, A. (1992). Determination of Na, K, Ca, Mg, Cu, Fe, Mn and total cationic milliequivalents in Spanish commercial honeys. *Journal of Apicultural Research, 31*, 65–69.

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.

Serra, J. (1989). Composición de la miel de eucalipto (*Eucalyptus* sp.) producida en España. *Annals de Bromatologia, 41*, 41–56.

Sharaf, M., Illman, D., & Kowalski, B. (1986). Preprocesing techniques. In *Chemometrics* (pp. 188–215). New York: Wiley.

Statgraphics (1991). *Statgraphics: reference manual*. Rockville, MD: STSC Inc.

Stein, K., & Umland, F. (1986). Trace analysis of lead, cadmium and manganese in honey and sugar. *Fresenius Journal of Analytical Chemistry, 323*, 176–177.

Vasconcelos, P., & Chaves, H. (1989). Classification of elementary wines of *Vitis Vinifera* varieties by pattern recognition of free amino acids profiles. *Journal of Agricultural and Food Chemistry, 37*, 931–937.

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. *Pattern Recognition, 8*, 127–139.

Wold, S., Albano, C., Dun, W. J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johanson, E., Lindberg, W., & Sjocaestrom, M. (1984). Multivariate data analysis in chemistry. In B. R. Kowalski (Ed.), *Chemometrics, mathematics and statistics in chemistry* (pp. 17–97). Dordrecht, The Netherlands: Riedel Publishing.