Characterization of Olive Oil Classes Using a ChemSensor and Pattern Recognition Techniques

F. Peña, S. Cárdenas, M. Gallego, and M. Valcárcel*

Department of Analytical Chemistry, Campus de Rabanales, University of Córdoba, E-14071, Córdoba, Spain

ABSTRACT: Classification is an important component of food quality assurance, as methods to guarantee authenticity of food products are widely demanded by food producers, processors, consumers, and regulatory bodies. The objective of this work was to develop a rapid classification method in order to discriminate virgin olive oil, olive oil, and "orujo" olive oil, the prices of which differ dramatically in the market on account of the high quality level of the former. For these purposes, new ChemSensor equipment that combines a headspace autosampler with a mass-selective detector and Pirouette data evaluation software was used. To take into account the large number of samples analyzed (50 samples repeated 10 times), as well as the wide interval of *m/z* ratios scanned (41–170), chemometric approaches were necessary. Cluster analysis, principal component analysis, K-nearest neighbors, and soft independent modeling of class analogy (SIMCA) were applied to model the different oil classes. The results indicated good classification and prediction abilities, with SIMCA affording the best results (*viz.* 97% specificity).

Paper no. J10275 in *JAOCS 79,* 1103–1108 (November 2002).

KEY WORDS: Chemsensor, classification, mass spectrometry, olive oils, pattern recognition.

Olive oil, and in particular virgin olive oil, is one of the basic components of the Mediterranean diet (1) and is classified into three main groups, namely, virgin olive oil, olive oil, and "orujo" olive oil. Virgin olive oil is extracted by purely mechanical means from sound, ripe fruits of the olive tree (*Olea europaea* L.). Olive oil is obtained by mixing refined olive oil and a variable amount of virgin olive oil (*ca.* 10–60%). Orujo is the residue remaining after virgin olive oil extraction with solvents. Orujo oil (pomace oil) is further processed to form refined orujo oil; by mixing the latter with small amounts of virgin olive oil (*ca.* 5–10%), orujo olive oil is produced, the third and lowest-quality type of olive oil.

Flavors and aroma of the oils are generated by a number of volatile constituents that are present at extremely low concentrations (2). Different isolation methods have been reported prior to the chromatographic separation/determination of volatile compounds, namely, direct injection, static and dynamic headspace, high-vacuum distillation, on-line LC–GC (3), and supercritical fluid extraction (4). The most recent methods have employed an electronic nose, a system that mimics human olfaction by combining the response of a set of chemical sensors, with partial specificity for the measurement of volatiles, and pattern recognition techniques for data interpretation (5–11). Various types of sensors, such as metal oxide semiconductors (5,7,11) and conducting polymers (6,8,10), have been used. The set of sensors of an electronic nose affords a large amount of information, and the processing of the data generated by the system is an essential part of the concept of electronic olfactometry.

Direct sampling-mass spectrometry (DS-MS) techniques are also related to the analysis of volatile compounds; they refer to the introduction of the analytes from a sample directly into a mass spectrometer using a simple interface with minimal sample preparation and no prior chromatographic separation (12). Thus, apart from avoiding typical problems in the employment of electronic noses—high sensitivity to some compounds, such as ethanol or water, and high cost of energy—DS-MS techniques possess several advantages, such as simplicity, real-time response, and high sample throughput. These techniques also provide a chemical fingerprint of the sample, which characterizes it and distinguishes it from other samples.

Chemometric techniques can also be used to process the data generated by the system; thus, multivariate calibration models, such as partial least squares (PLS) regression, tri-PLS, and parallel factor analysis (13), are useful for the analysis of sample mixtures containing analytes with similar mass spectra. Vegetable oil classification has been carried out by dynamic headspace GC using ANOVA (14), by ^{13}C NMR using PLS and principal component analysis (PCA) (15), or K-nearest neighbors (KNN), soft independent modeling of class analogy (SIMCA), and linear discriminant analysis (11).

Recently, Agilent Technologies (Palo Alto, CA) commercialized a new instrument, the ChemSensor 4440, which consists of a headspace autosampler with a mass spectrometer for screening purposes. Few applications of this instrument have been developed to date; the basic operating principles and fields of use of the ChemSensor 4440 have been evaluated, and examples of the possibilities of use of the instrument in process and quality control in the pharmaceutical, food, and cosmetic industries have been given (16,17). Recently, this technology was successfully applied to the detection of adulterants in olive oil (18). In the present work, a novel application of this instrument is proposed in order to classify three different edible olive oils, namely, virgin olive oil, olive oil, and orujo olive oil. Edible oils were automatically carried from the autosampler to the heating unit and 3 mL of the headspace generated was transferred by a helium carrier

*To whom correspondence should be addressed at Department of Analytical Chemistry, Edif. C-3 (anexo), Campus de Rabanales, 14071-Córdoba, Spain. E-mail: qa1meobj@uco.es stream into the mass spectrometer. Taking into account the wide *m/z* range selected (41–170) and the numerous samples analyzed (500), chemometric treatment was necessary. Thus, four pattern recognition techniques [cluster analysis (CA), PCA, KNN, and SIMCA] included in the Pirouette data evaluation software of the instrument provided by Infometrix Inc. (Woodinville, WA) were applied.

EXPERIMENTAL PROCEDURES

***Apparatus*.** Oil analyses were performed with a ChemSensor 4440B (Agilent Technologies) system, which comprises two modules. The first one is a 44-space autosampler for headspace vials that includes a robotic arm and a headspace generation unit with two parts: an oven to heat the samples inside the vials and form the headspace, and a six-port injection valve with a 3-mL loop. Helium (5.0 grade purity, Air Liquide, Seville, Spain), regulated by a digital pressure and flow controller, was used for both pressurizing the vial (18.0 psi of flow pressure) and carrying the headspace formed to the detector (2.0 psi of flow pressure). Every tubing of this unit, together with a transfer line connected to the heated interface of the detector, is passivated with Silicosteel. The second module is a quadrupole 5973 mass spectrometer detector, operated in full scan mode with a mass range between *m/z* 41 and 170. EI ionization was used with an ionization energy of 70 eV. The transfer line, source, and quadrupole temperatures were maintained at 120, 200, and 120°C, respectively. Total ion current chromatograms were acquired and processed using G1701BA Standalone data analysis software (Infometrix) on a Pentium II computer that also controlled the whole system.

Ten-milliliter glass flat-bottomed vials for headspace analysis with 20-mm polytetrafluoroethylene/silicone septa (Supelco, Madrid, Spain) were also employed.

Oil samples. Twenty-nine samples of virgin olive oil (VOO), 16 samples of olive oil (OO), and 5 samples of orujo olive oil (OOO) were obtained from different local markets in Spain. All samples were from Andalusian cultivars (south of Spain) and corresponded to Picual and Hojiblanca ripe fruit varieties. Ten aliquots of each sample were analyzed, so 500 analyses were performed; a small number of aliquots showed poor reproducibility. In order to prevent volatile losses or contamination of the volatile fraction, all samples were stored in a cold, dark place.

Procedure. Aliquots of 5.0 mL of each oil sample were added to the 10-mL headspace vials and placed into the autosampler. The robotic arm took each vial from the 44-space carousel and placed it into the oven; the sample was then heated at 90ºC for 30 min in order to enrich and equilibrate the gaseous phase in the volatile compounds of the oil sample. Afterward (Fig. 1A), a needle connected to the injection valve (IV) entered the vial and a helium line pressurized the headspace for 12 s; then, by opening a vent valve, and due to the different pressure inside the vial and at the end of the tubing (atmospheric pressure), volatile compounds were driven out of the vial *via* the needle, filling the 3-mL loop of the IV, previously heated at 110ºC, and being released to the atmosphere during 3 s. In a second step (Fig. 1B), the IV was switched and the helium stream transported the loop contents to the mass spectrometer.

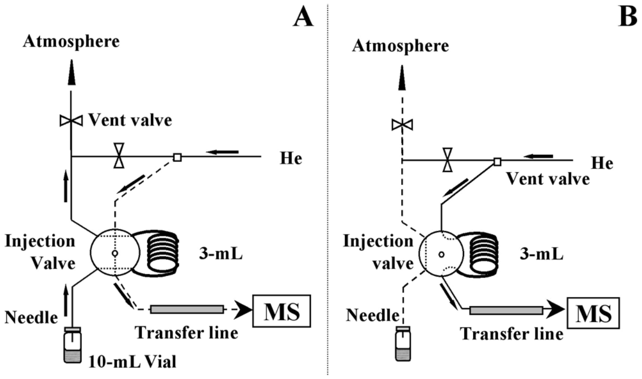

FIG. 1. Diagram of pressurizing/venting (A) and injection (B) positions of the headspace generation unit.

In this scheme, there is no chromatographic separation, so the volatiles arrive at the detector at the same time, providing a total ion current chromatogram, which can be considered a chemical fingerprint of the oil sample, called a volatiles profile, and which can be used for classification purposes.

Chemometric procedures. The data set consisted of a 500 × 130 data matrix in which the rows corresponded to the different oil samples (290 of virgin olive oil, 160 of olive oil, and 50 of orujo olive oil) and the columns to the 130 masses scanned by the MS detector, from *m/z* 41 to 170. Signals obtained from the detector were treated first by two unsupervised techniques (CA and PCA), in order to find internal structures or clustering of data, and later by two supervised techniques (KNN and SIMCA), in order to obtain adequate classification procedures, all of which are explained below. For unsupervised procedures, a 95% confidence level was fixed.

All chemometric analyses were performed by means of the statistical software package "Pirouette: Multivariate Data Analysis," developed by Infometrix Inc.

RESULTS AND DISCUSSION

CA. CA is a complementary technique to describe the structure of a data table and search for natural groupings among the samples, commonly applied before other multivariate procedures owing to its unsupervised nature. In this work, similarities between the samples were calculated on the basis of the Euclidean distance, while a hierarchical agglomerative procedure with complete linkage was used to establish clusters. The proximity between two sample responses (*viz.*, *j* and *j*′, max) can be represented by plotting either the multivariate distance *d* or using a similarity index (*S*) to normalize (19), where:

$$S = 1 - \frac{d_{jj'}}{d_{\max}} \qquad [1]$$

The similarity index lies between 0 and 1, taking a value of zero for the most different sample responses and a value of unity for identical sample responses. CA was applied to the raw data, but neither preprocessing techniques nor transformations improved the results. As can be seen in Figure 2, at an *S* of 0.73, seven clusters (A–G) were established, but almost all of them overlapped with adjacent clusters; no clear separation was evident to help in the classification. Due to the large number of samples analyzed (500 samples), and the larger number of samples of one type (290 VOO) than of the other two types (160 OO and 50 OOO), there is a certain overlap between the three classes studied. Only two clusters are composed of samples belonging just to the VOO class, the most abundant kind of oil (clusters C and D). On the other hand, of the samples in cluster A, 31.7% are VOO, 55.2% are OO, and 13.1% are OOO samples, whereas in cluster B, 52.9% are VOO and 47.1% are OO samples. As indicated before, clusters C and D are 100% composed of VOO samples, and clusters E to G, containing 7, 4, and 1 samples, respectively, are composed of misgrouped samples. Such an irregular clustering can be explained on the basis of the larger number of VOO samples analyzed in comparison with the OOO samples. Therefore, no clear separation is achieved employing clustering techniques.

PCA. PCA is used to reduce the dimensionality of the data matrix, retaining the maximal amount possible of variability present in the original data. Essentially, the original *m* variables—in our case, *m/z* ratios from the MS detector—are linearly combined to form *F* new variables called factors or principal components (19). PCA represents the original data matrix ($X_{n\times m}$), where *n* corresponds to the number of samples and *m* to the number of raw variables (from *m/z* 41 to 170), respectively, as a product of two matrices: the scores matrix ($S_{n\times F}$) and the loadings matrix ($L_{F\times m}$), plus a table of residuals ($E_{n\times m}$). This corresponds to projecting the *X* matrix down on a few-dimensional (*F*) space. When the number of factors is small compared with *m*, PCA provides a considerable simplification and reduction of the data matrix. The scores are the values of the samples represented in the new *F*-dimensional space, and the loadings are the coefficients of the combined *m* original variables. PCA was performed on the mean-centered data: from the loadings of the original variables in the three first principal components, *m/z* ratio 44 is the dominant variable in the first principal component, which represents 97.6% of the total variability; *m/z* ratios 43 and 58 dominate the second principal component, which represents 2.1%, and *m/z* ratios 45 and 46 dominate the third principal component, which represents 0.2%. As can be seen in Figure 3, where the scores of each oil sample are examined in a 3-D plot of the first three principal components, discrimination between VOO and OO samples is not clear; however, there is a correct classification among OOO samples, which are perfectly grouped. The OO samples are distributed in five subgroups as a consequence of the different percentage of VOO they contain (between 10 and 60%). The two subgroups closer to the OOO samples consisted of those samples with the lowest percentage of VOO; their volatile profiles were more nearly similar to the OOO samples, for which the percentage of VOO is *ca.* 7%. In addition, an irregular and dispersed distribution of VOO samples was obtained, which can be attributed to the large number of samples analyzed (29 out of 50 samples) because of the higher availability of that sample in the studied region. These samples are directly extracted, by purely mechanical means, from the olives, and their volatile fractions vary with both the variety of olives (*viz.,* Hojiblanca, Aloreña, Picual, Arbequina, etc.) and the acidity (*viz.,* 1–3°), among other characteristics.

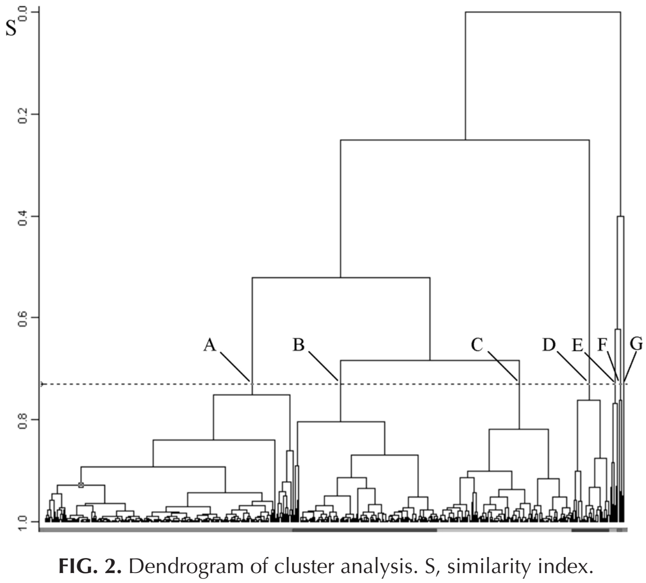

FIG. 2. Dendrogram of cluster analysis. S, similarity index.

The study of the latent structures residing in the data set, obtained after the application of the two unsupervised procedures (CA and PCA), revealed great similarities among the volatiles composition of the three types of oil samples studied, especially between VOO and OO, and the difficulty of classifying them on the basis of only their volatiles profiles. There is clear evidence of the different characteristics of OOO samples that permits their discrimination from other samples employing PCA. A larger amount of OOO samples should be studied in order to confirm these results, but in the geographical region studied, few manufacturers produce this kind of oil.

KNN. KNN is a discriminant, nonparametric technique (11), based on the distance between objects in a space of dimension equal to the number of variables explored. The class to which the sample is assigned is that of the samples of the training set that are closest to it. Only the K closest objects are used to make the assignment. From a mathematical point of view, it is a simple method, free from statistical assumptions. The distance criterion used in the present work was Euclidean distance. Initially, a classification model was constructed in which all samples were used as the training set. In a second stage, to validate the classification model thus obtained and its stability in predicting, a cross-validation step was performed with five cancellation groups (the samples were randomly divided into five groups, each of them containing 20% of the total), four of which were used as the training set and the fifth as the prediction set. To perform this cross-validation procedure, the same process was repeated five times with the five different training and prediction sets,

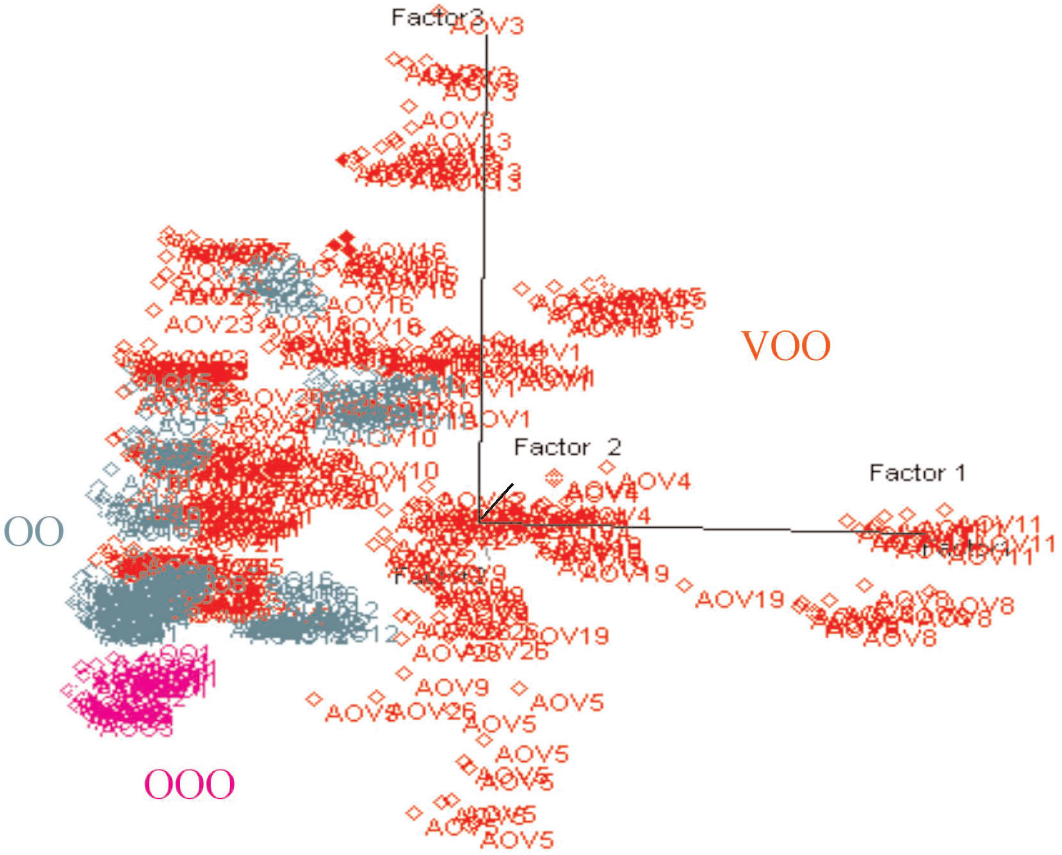

FIG. 3. Plot of oil sample scores for principal component analysis model. Virgin olive oil (VOO, in red), olive oil (OO, in green), and orujo olive oil (OOO, in pink).

ensuring that all the samples were included at least once in the prediction set. The success of this classification system was expressed as classification ability (percentage of members of the training set correctly classified) and prediction ability (percentage of the test set members adequately classified by using the rules developed in the training step). The term "specificity" refers to the percentage of samples that, belonging to a different class, are recognized as being foreign to the model; the specificity for each category was related to the other two classes, so that specificity for VOO samples was evaluated in relation with OO or with OOO classes. KNN was applied, with raw and normalized data (raw data divided by their maximum value), for the classification of VOO, OO, and OOO samples. The value of K was selected by optimization, determining the classification ability and the number of misclassifications, with K values between 1 and 10; the best results were achieved using K = 1 (fewer number of misclassifications obtained). Therefore, K = 1 was selected for the application of KNN. Under these conditions, the results for both raw and normalized data were as summarized in Table 1. According to these data, good classification and prediction percentages were obtained, indicating that the mathematical

TABLE 1
Percentages of Classification and Prediction Abilities for KNN and SIMCA

		KNN[a]				SIMCA[a]			
		Raw data		Normalized data		Raw data		Normalized data	
		1	2	1	2	1	2	1	2
Classification	VOO	92.8	93.1	95.8	96.2	96.7	62.4	96.3	74.5
	OO	87.8	90.0	90.2	90.0	99.8	68.1	98.5	57.0
	OOO	70.0	70.0	84.2	90.0	100.0	100.0	100.0	60.0
Prediction	VOO	90.7		94.8		94.1		91.7	
	OO	91.9		92.5		92.5		93.1	
	OOO	63.3		93.3		63.3		60.0	

[a] 1 corresponds to cross-validation with five cancellation groups; 2 corresponds to correct classification with all samples in the training set. KNN, K-nearest neighbors; SIMCA, soft independent modeling of class analogy; VOO, virgin olive oil; OO, olive oil; OOO, orujo olive oil.

TABLE 2
Specificity for Each Type of Olive Oil Using KNN and SIMCA[a]

		VOO		OO		OOO	
	Specificity	KNN	SIMCA	KNN	SIMCA	KNN	SIMCA
Raw data	VOO			91.9	99.8	96.7	100.0
	OO	93.4	98.4			73.3	100.0
	OOO	99.5	100.0	96.5	100.0		
Normalized data	VOO			93.8	99.4	95.0	100.0
	OO	95.6	97.4			85.0	100.0
	OOO	99.0	100.0	94.2	100.0		

[a] See Table 1 for abbreviations.

models generated can correctly classify and discriminate among the different types of oil samples. Classification ability did not change significantly when the training set was composed of only 80% of the samples (column 1, in Table 1) versus 100% of them (column 2, in Table 1), except for OOO samples with normalized data. However, normalization of the data improved the overall classification ability for the OOO class owing to the smaller number of samples available. Similar values were obtained for classification and prediction abilities, indicating that the model is fairly stable. In considering those results, KNN provides an adequate model to classify the different types of olive oil on the basis of their volatiles profiles. Specificity for KNN was studied using 80% of the samples as a training set (cross-validation). The results are listed in Table 2, where raw and normalized data are also compared. In general, good percentages for every studied class were obtained; the poorest results were achieved for the OO class (73.3 or 85.0%) in relation to OOO. Normalization of data did not help in the classification task, as it provided similar results as the raw data.

According to the results listed in Tables 1 and 2, the proposed KNN model provides classification and prediction abilities that make it adequate for the proposed task of recognizing the different types of oil samples.

SIMCA. The last supervised pattern recognition technique applied to the data is the most complicated as it is based on the principal components of each category and critical distances with probabilistic significance (19). It is a class-modeling technique that builds frontiers between each class and the rest of the universe; with this technique the classification rule for a given class is a class-box that envelops the position of the class in a pattern space, so that an object is assigned to a class if it is situated inside the boundaries of only one class-box and considered to be an outlier for that class if it falls outside the class-box (11). SIMCA constructs an independent model for each class by PCA; the number of principal com-

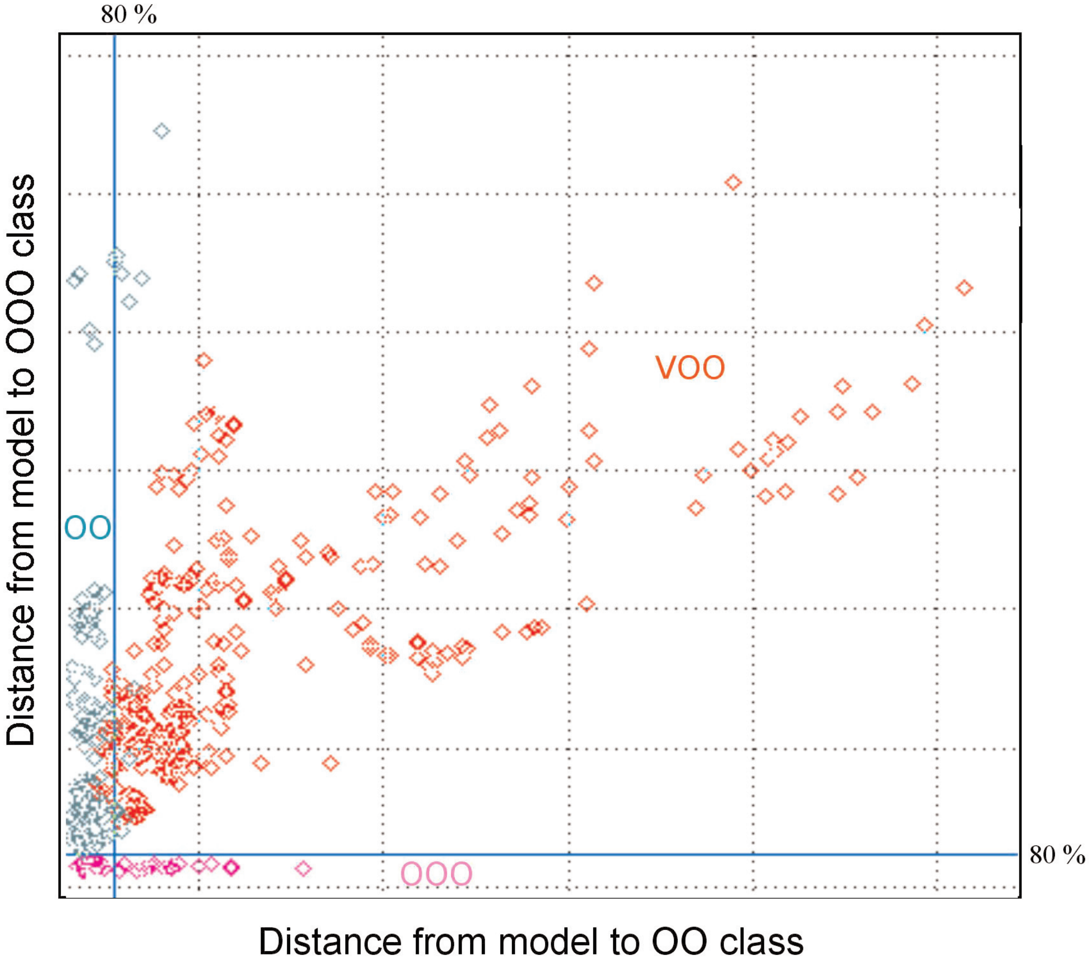

FIG. 4. Coomans plot, for soft independent modeling of class analogy, of oil samples. For abbreviations and key see Figure 3.

ponents used for each class may be preset or may be selected in such a way that they explain a given percentage of the variance of the data. In this way, a closed space is constructed on the basis of a critical distance. Each object considered is assigned to one category according to its Euclidean distance from the model. The same concept of specificity, defined above, can be associated with this chemometric treatment. Apart from a correct classification using all samples as the training set, a cross-validation procedure in five steps, as described for KNN, was used. Again, models were obtained for both raw and normalized data. In both cases, the model created afforded 12 components for the two first categories (VOO and OO) and nine components for the third category (OOO), so that a great reduction in the number of variables was made with almost no loss of information. The results were studied in terms of classification and prediction abilities and are summarized in Table 1. The classification model developed by SIMCA in the validation step produced good results for classification abilities (percentages higher than 96% both for raw and normalized data). Poor results were achieved for correct classification except for raw data of OOO samples (100%); for prediction abilities, similar results were obtained, although OOO samples again provided the poorest percentages. In addition, a Coomans plot was used (Fig. 4) to evaluate the category classification of oil samples; the axes of the plot represent the distances of the samples from the models to the OO and OOO classes, respectively. The majority of VOO samples were obviously considered outliers, but there were some that were included in the OO region; this overlap was already discussed in relation to the CA and PCA models. OO and OOO samples were located in their own region, and some OOO samples appearing in the region related to both classes. Specificity results for SIMCA are listed in Table 2; better results were obtained than with KNN (more than 97% specificity with both raw and normalized data). This advantage can be explained on the basis of the different way in which the classification task was carried out. As seen in Tables 1 and 2 and Figure 4, SIMCA was the technique that offered the best results in the classification task, with percentages of specificity of 100.0% for VOO and OO classes vs. the OOO class.

ACKNOWLEDGMENTS

This work was supported by grant BQU 2001-1815 from Spain's Dirección General de Investigación, Ciencia y Tecnología.

REFERENCES

1. *Manual del Aceite de Oliva*, www.aceitedeoliva.com.
2. Warner, K., and T. Nelsen, AOCS Collaborative Study on Sensory and Volatile Compound Analyses of Vegetable Oils, *J. Am. Oil Chem. Soc. 73*:157–162 (1996).
3. Cert, A., W. Moreda, and M.C. Pérez-Camino, Chromatographic Analysis of Minor Constituents in Vegetable Oils, *J. Chromatogr. A 881*:131–148 (2000).
4. Morales, M.T., A.J. Berry, P.S. McIntyre, and R. Aparicio, Tentative Analysis of Virgin Olive Oil Aroma by Supercritical Fluid Extraction–High Resolution Gas Chromatography–Mass Spectrometry, *Ibid. 819*:267–275 (1998).
5. Gonzalez-Martín, Y., J.L. Pérez-Pavón, B. Moreno Cordero, and C. Garcia Pinto, Classification of Vegetable Oils by Linear Discriminant Analysis of Electronic Nose Data, *Anal. Chim. Acta 384*:83–94 (1999).
6. Stella, R., J.N. Barisci, G. Serra, G.G. Wallace, and D. De Rossi, Characterization of Olive Oil by an Electronic Nose Based on Conducting Polymer Sensors, *Sens. Actuat. B 63*:1–9 (2000).
7. Capone, S., P. Siciliano, F. Quaranta, R. Rella, M. Epifani, and L. Vasanelli, Analysis of Vapors and Foods by Means of an Electronic Nose Based on a Sol-Gel Metal Oxide Sensors Array, *Ibid. 69*:230–235 (2000).
8. Guadarrama, A., M.L. Rodríguez-Méndez, J.A. de Saja, J.L. Ríos, and J.M. Olías, Array of Sensors Based on Conducting Polymers for the Quality Control of the Aroma of the Virgin Olive Oil, *Ibid. 69*:276–282 (2000).
9. Koprivnjak, O., G. Procida, and T. Zelinotti, Changes in the Volatile Components of Virgin Olive Oil During Fruit Storage in Aqueous Media, *Food Chem. 70*:377–384 (2000).
10. Guadarrama, A., M.L. Rodríguez-Méndez, C. Sanz, J.L. Ríos, and J.A. de Saja, Electronic Nose Based on Conducting Polymers for the Quality Control of the Olive Oil Aroma. Discrimination of Quality, Variety of Olive and Geographic Origin, *Anal. Chim. Acta 432*:283–292 (2001).
11. Gonzalez-Martín, Y., M.C. Cerrato Oliveros, J.L. Pérez-Pavón, C. Garcia Pinto, and B. Moreno Cordero, Electronic Nose Based on Metal Oxide Semiconductor Sensors and Pattern Recognition Techniques: Characterization of Vegetable Oils, *Ibid. 449*:69–80 (2001).
12. Wise, M.B., and M.R. Guerin, Direct Sampling MS for Environmental Screening, *Anal. Chem. News Features 1*:26A–32A (1997).
13. Gardner, W.P., R.E. Shaffer, J.E. Girard, and J.H. Callahan, Application of Quantitative Chemometric Analysis Techniques to Direct Sampling Mass Spectrometry, *Anal. Chem. 73*:596–605 (2001).
14. Morales, M.T., R. Aparicio, and J.J. Ríos, Dynamic Headspace Gas Chromatographic Method for Determining Volatiles in Virgin Olive Oil, *J. Chromatogr. A 668*:455–462 (1994).
15. Shaw, A.D., A. Di Camillo, G. Vlahov, A. Jones, G. Bianchi, J. Rowland, and D.B. Kell, Discrimination of the Variety and Region of Origin of Extra Virgin Olive Oils Using ^{13}C NMR and Multivariate Calibration with Variable Reduction, *Anal. Chim. Acta 348*:357–374 (1997).
16. Anon., Quality Control with the ChemSensor, *GIT Labor Fachz. 44*:228–230 (2000).
17. Filadeau, Y., Electronic Nose with MS Detector for Rapid Measurement of the Chemical Integrity of Samples, *Spectral Analysis 29*:37–38 (2000).
18. Marcos Lorenzo, I., J.L. Perez-Pavón, M.E. Fernández Laespada, C. García Pinto, and B. Moreno Cordero, Detection of Adulterants in Olive Oil by Headspace-Mass Spectrometry, *J. Chromatogr. A 945*:221–230 (2002).
19. Gardner, J.W., and P.N. Bartlet, *Electronic Noses: Principles and Applications*, Oxford University Press, New York, 1999, pp. 151–155.

[Received March 19, 2002; accepted July 12, 2002]