JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

J. Agric. Food Chem. **2002,** *50,* 413−418 **413**

Chemical Authentication of Extra Virgin Olive Oil Varieties by Supervised Chemometric Procedures

REMO BUCCI,[†] ANDREA D. MAGRÍ,[†] ANTONIO L. MAGRÍ,[†] DOMENICO MARINI,[‡] AND FEDERICO MARINI*,[†]

Dipartimento di Chimica, Università degli Studi "La Sapienza", P.le A. Moro 5, I-00185 Roma, Italy, and Dipartimento delle Dogane, DCAMLC Div.VI, Via M. Carucci 71, I-00143 Roma, Italy

This work has focused on discriminating extra virgin olive oils from Sabina (Lazio, Italy) by olive fruit variety (cultivar). A set of oils from five of the most widespread cultivars (Carboncella, Frantoio, Leccino, Moraiolo, and Pendolino) in this geographical area was analyzed for chemical composition using only the Official Analytical Methods, recognized for the quality control and commercial classification of this product. The obtained data set was converted into a computer-compatible format, and principal component analysis (PCA) and a method based on the Fisher *F* ratio were used to reduce the number of variables without a significant loss of chemical information. Then, to differentiate these samples, two supervised chemometric procedures were applied to process the experimental data: linear discriminant analysis (LDA) and artificial neural network (ANN) using the back-propagation algorithm. It was found that both of these techniques were able to generalize and correctly predict all of the samples in the test set. However, these results were obtained using 10 variables for LDA and 6 (the major fatty acid percentages, determined by a single gas chromatogram) for ANN, which, in this case, appears to provide a better prediction ability and a simpler chemical analysis. Finally, it is pointed out that, to achieve the correct authentication of all samples, the selected training set must be representative of the whole data set.

KEYWORDS: Olive oil; pattern recognition; linear discriminant analysis; artificial neural network

INTRODUCTION

Although it is commonly accepted that safety of use and quality are prerequisites for foodstuffs, costs, economic interests, and experimental difficulties have up to now curbed their systematic certification. This problem now appears more definite and important as, according to the new European Community agricultural policies and subsequent international agreements (e.g., Schengen Treaty), commodities having the same commercial denomination but different quality and production costs may enter the market.

Therefore, to protect some productive identities which would otherwise be at risk to competition, norms have been introduced to classify goods in a more effective way, to differentiate them on the basis of their principal peculiarities.

Extra virgin olive oil is one product that—for its alimentary importance and wide distribution, especially in Mediterranean countries—is affected by this situation. The present community provisions make reference to the determination of some chemical indices to check the genuineness of olive oils, but these indices, separately considered, are unable to supply complete information on the quality of the product. To assess oil quality, which can depend on a given geographical origin, extraction method, or olive fruit variety (cultivar), analytical data have indeed to be processed as a whole, using mathematical and statistical techniques. Accordingly, many scientific approaches have been undertaken to study useful procedures and indices. Several of these studies, however, aimed at finding new indices, often determined by using expensive instrumentation that is unavailable in most of quality control laboratories (*1*−*7*). Therefore, many chemical operators may have difficulties in using these indices, also because it is difficult to find in the literature a suitable data set to increase their own in order to perform a meaningful statistical analysis.

Continuing our research activity in the field of value-added foodstuff quality control and certification, especially on oil (*8*−*12*), in this work we focused on authenticating extra virgin olive oils from Sabina (Lazio, Italy) by their variety. To overcome the difficulties previously stated, we have chosen the variables necessary for this discrimination from those determinable according to the Official Analytical Methods enacted by the European Union for the genuineness control and the commercial classification of this product. These indices, some of which have been already successfully used for the geographical authentication of extra virgin olive oils (*1*, *13*, *14*), can be easily

* Author to whom correspondence should be addressed (telephone +39 06 49913371; fax +39 06 4457050; e-mail fmmonet@hotmail.com).
[†] Università degli Studi "La Sapienza".
[‡] DCAMLC Div. VI.

10.1021/jf010696v CCC: $22.00
Published on Web 01/04/2002

determined in every chemical laboratory, and the results of the analyses from a great number of samples can be found in the literature.

On the basis of these considerations, a set of oils from five of the most widespread cultivars (Carboncella, Frantoio, Leccino, Pendolino, and Moraiolo) in Sabina was analyzed for their chemical composition (acidity, peroxide value, UV absorbance, major fatty acid, triglycerides, and sterol content). To differentiate these samples, we processed the obtained experimental data applying the following supervised chemometric procedures: linear discriminant analysis (LDA) and back-propagation artificial neural network (BP-ANN) (*15*−*18*).

MATERIALS AND METHODS

Data Set. The data set for the statistical analysis was made up of the results of the chemical analyses performed upon 153 extra virgin olive oils (years of harvesting 1997−1999), but only the samples produced in 1999 (53 oils) were analyzed in our laboratory; all other data were taken from Customs Chemical Laboratories archives. Chemical analyses have been performed in the two laboratories in a consistent way, according to reference methods (*19*). To authenticate olive oil varieties, all of the samples (extra virgin oils) were obtained according to the same extracting procedure (cold pressing), using monocultivar fruits produced in a well-defined area (Sabina, Lazio, Italy) and stored in dark bottles at 4 °C until chemical analyses were performed.

In particular, the following five cultivars, all representative of this geographical area, were selected (the numbers of samples analyzed in our laboratory and in Customs Laboratories is reported in parentheses): Carboncella (8 + 20), Frantoio (16 + 21), Leccino (15 + 22), Moraiolo (7 + 26), and Pendolino (6 + 12).

Procedures. All of the oil samples were analyzed for acidity (percent oleic acid), K_{270} (UV specific extinction coefficient at $\lambda = 270$ nm), peroxide value, major fatty acids composition (palmitic, palmitoleic, stearic, oleic, linoleic, and linolenic), trilinolein content, and sterol composition (cholesterol, brassicasterol, 24-methylenecholesterol, campesterol, campestanol, stigmasterol, $\Delta^{5,23}$-stigmastadienol, clerosterol, β-sitosterol, sitostanol, Δ^5-avenasterol, Δ^7-campesterol, $\Delta^{5,24}$-stigmastadienol, Δ^7-stigmasterol, and Δ^7-avenasterol).

The analyses were performed according to the respective Official Methods (*19*), the repeatability of which in our experimental conditions was checked by six replicates on a randomly chosen sample. The chemical analyses of all the other samples were performed at least twice.

Apparatus. UV spectra for K_{270} determination were recorded using a Perkin-Elmer 320 UV−visible spectrophotometer furnished with 1 cm quartz cells.

GC data were obtained by using a Fisons HRGC Mega II Series model 8560 equipped with a split/splitless injector and an FID detector and connected to a computer provided with the analytical program Chrom-Card. The chromatographic columns used were a Chrompack CP-Sil 88 capillary column ($l = 50$ m; i.d. = 0.25 mm; film thickness = 0.20 μm) for fatty acid methyl esters and an Alltech SE-54 capillary column ($l = 30$ m; i.d. = 0.25 mm; film thickness = 0.25 μm) for sterol silyl ethers analyses. Helium at flow rate of 2 mL min^{-1} was used as carrier gas.

Merck Silica Gel 60 TLC plates (200 × 200 mm; thickness = 25 μm) were used to separate the sterol fraction from the unsaponifiable fraction.

HPLC of triacylglycerides was performed on a ThermoQuest HPLC model Spectra Series, equipped with two Supelco Sil LC-18 columns in series ($l = 15$ cm; i.d. = 4.5 mm; particle size = 5 μm) and a Varian RI4 refractive index detector and connected to a computer provided with the analytical program Chrom-Card. A 60:40 (v/v) acetone/acetonitrile mixture at flow rate of 0.8 mL min^{-1} was used as the mobile phase.

All of the reagents were of the required purity grade.

Statistical Data Analysis. The analytical data were arranged in a matrix to perform the statistical analysis for the variety authentication.

It is well-known (*20*−*23*) that redundant and/or less discriminating variables contain a lot of noise, which affects the chemometric predictive ability, so at first basic statistics, either Anova (by means of the Fisher F ratio), factor analysis, and/or principal component analysis (PCA), are normally employed to reduce the number of variables to be included in the mathematical model. In particular, PCA (*24*−*27*) bases itself on projecting the data matrix into an orthogonal subspace, the axes (principal component or factors) of which are linear combinations of the original variables, such that the first principal component accounts for the largest amount of the variance originally present in the data set, the second for the largest amount of the residual variance unexplained by the first, and so on until the total sample variance is combined into component groups. If each factor can be identified with a corresponding chemical index, it is usually possible to achieve a simplification of the chemical analyses, reducing times and costs significantly. Simplification is not always possible using PCA, where Varimax rotation of factors can only sometimes result in an effective matching of rotated principal components and experimental indices. However, the use of a variable reduction method based on the Fisher F ratio allows the retention of the more discriminating indices in the model. In fact, the Fisher F ratio is defined as the ratio of between-groups variance to within-group variance, according to the formula

$$F(y) = \frac{\{1\}/\{(g-1)\}}{\{1\}/\{(n-g)\}} \frac{\sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2} \quad (1)$$

If F has a value of <1, the variable should be discarded as it represents a hindrance to correct variety discrimination. It should be noted, however, that several other factors (e.g., stage of ripeness, year of harvesting, and region of origin) could affect values of this ratio, so it is advisable to choose a threshold value >1 and proceed downward toward $F = 1$ to find the best model (*5*).

When the classifications of some samples are known a priori (standards)—as in our case—the final classification of samples in the groups (pattern recognition) is usually carried out using supervised techniques (LDA and BP-ANN). In this case, the entire group of analytical data, collected from the standard samples, is divided into two sets: the training and the test set. The first is employed to build a classification rule, which afterward allows the attribution of unknown samples, whereas the second is often used to validate the predictive ability of the optimized mathematical model. Whatever the technique chosen, the classification rule that provides the better prediction is "a sample should be assigned to the class for which it has the largest posterior probability" (Bayes' rule) (*28*). The posterior probability, for each class, can be computed according to Bayes' theorem

$$P(G_i|\mathbf{x},H_0) = \frac{P_0(G_i|H_0)p(\mathbf{x}|G_i,H_0)}{\sum_i P_0(G_i|H_0)p(\mathbf{x}|G_i,H_0)} \quad (2)$$

where $P_0(G_i|H_0)P(G_i|\mathbf{x},H_0)$ is the posterior probability that a sample described by the random vector $\mathbf{x}$ belongs to group G_i, $P_0(G_i|H_0)$ is the a priori probability of observing a sample from group G_i (**prior**), and $p(\mathbf{x}|G_i,H_0)$ is the conditional probability of observing a random vector $\mathbf{x}$ for samples belonging to group G_i (**likelihood**) (*29*).

LDA (*30*−*33*) assumes that conditional probabilities are normally distributed with the same variance−covariance matrix for each group, and because probabilities are positive, taking the natural logarithm of posterior probabilities gives for each group a classification function that is linear in the original variables, hence, the name of this method. In this case, the above-mentioned Bayes' rule becomes "a sample should be assigned to the class whose classification function is largest".

BP-ANN (*34*−*39*) constitutes a valid alternative to classical pattern recognition techniques, as the outputs of a well-trained network can be interpreted as posterior probabilities; here Softmax Output should be used because network outputs should meet the mathematical

Table 1. Mean Values and Standard Deviations of the Chemical Indices within Each Olive Cultivar and the Whole Data Set

variable	Carboncella	Frantoio	Leccino	Moraiolo	Pendolino	cumulative
acidity (g/100 g)	$0.6_4 \pm 0.1_5$	$0.3_7 \pm 0.1_9$	$0.3_7 \pm 0.1_6$	$0.6_4 \pm 0.1_6$	$0.2_5 \pm 0.1_1$	$0.4_6 \pm 0.2_2$
palmitic (%)	$12.8_4 \pm 0.4_8$	$13._8 \pm 1._2$	$13._9 \pm 1._4$	$11.3_0 \pm 0.4_5$	$14.0_4 \pm 0.3_4$	$13._1 \pm 1._4$
palmitoleic (%)	$0.66_2 \pm 0.09_5$	$1.2_7 \pm 0.2_3$	$1.3_3 \pm 0.1_5$	$0.63_7 \pm 0.08_5$	$1.3_8 \pm 0.1_4$	$1.0_5 \pm 0.3_6$
stearic (%)	$1.9_6 \pm 0.1_6$	$1.7_9 \pm 0.1_9$	$1.6_9 \pm 0.1_6$	$2.0_5 \pm 0.1_3$	$2.2_3 \pm 0.1_5$	$1.9_1 \pm 0.2_4$
oleic (%)	$75.7_4 \pm 0.7_5$	$76._1 \pm 3._3$	$74._8 \pm 1._3$	$77.2_1 \pm 0.6_7$	$74.2_7 \pm 0.4_0$	$75._8 \pm 2._0$
linoleic (%)	$7.2_3 \pm 0.4_0$	$6._7 \pm 1._1$	$6.0_6 \pm 0.1_9$	$6.9_4 \pm 0.1_3$	$6.5_1 \pm 0.2_7$	$6.6_8 \pm 0.6_1$
linolenic (%)	$0.80_7 \pm 0.06_2$	$0.5_7 \pm 0.1_6$	$0.6_7 \pm 0.1_2$	$0.57_9 \pm 0.05_0$	$0.49_6 \pm 0.05_6$	$0.6_3 \pm 0.1_5$
cholesterol (%)	$0.3_0 \pm 0.1_0$	$0.34_5 \pm 0.07_1$	$0.38_4 \pm 0.09_3$	$0.36_5 \pm 0.07_0$	$0.36_3 \pm 0.06_9$	$0.35_3 \pm 0.08_6$
campesterol (%)	$2.8_0 \pm 0.2_5$	$3.1_3 \pm 0.3_2$	$3.1_7 \pm 0.4_7$	$3.4_0 \pm 0.1_4$	$3.2_7 \pm 0.1_7$	$3.1_5 \pm 0.3_6$
stigmasterol (%)	$1.2_5 \pm 0.2_3$	$1.2_2 \pm 0.3_9$	$1.2_9 \pm 0.3_7$	$1.6_3 \pm 0.2_0$	$0.9_7 \pm 0.2_2$	$1.3_0 \pm 0.3_6$
clerosterol (%)	$0.8_6 \pm 0.1_1$	$0.8_1 \pm 0.1_8$	$0.9_0 \pm 0.2_2$	$0.8_9 \pm 0.1_4$	$0.9_8 \pm 0.2_4$	$0.8_8 \pm 0.1_8$
β-sistosterol (%)	$80._0 \pm 3._8$	$80._8 \pm 2._4$	$81._0 \pm 2._3$	$83._1 \pm 2._8$	$82._3 \pm 3._5$	$81._4 \pm 3._1$
sitostanol (%)	$0.7_6 \pm 0.1_9$	$1.0_5 \pm 0.3_2$	$1.1_4 \pm 0.3_6$	$1.2_5 \pm 0.2_9$	$1.0_5 \pm 0.2_7$	$1.0_6 \pm 0.3_4$
$\Delta^{5,24}$-stigmastadienol (%)	$0.6_4 \pm 0.1_5$	$0.6_3 \pm 0.2_3$	$0.5_8 \pm 0.1_7$	$0.5_7 \pm 0.2_1$	$0.5_4 \pm 0.2_6$	$0.6_0 \pm 0.2_0$
Δ^7-sigmasterol (%)	$0.33_2 \pm 0.07_0$	$0.21_0 \pm 0.09_4$	$0.17_4 \pm 0.06_4$	$0.23_6 \pm 0.06_6$	$0.15_1 \pm 0.04_4$	$0.22_2 \pm 0.09_3$
Δ^7-avenasterol (%)	$0.7_9 \pm 0.2_2$	$0.6_0 \pm 0.2_7$	$0.4_8 \pm 0.1_5$	$0.4_7 \pm 0.2_1$	$0.40_1 \pm 0.09_9$	$0.5_6 \pm 0.2_4$
LLL (%)	$0.23_1 \pm 0.05_1$	$0.23_8 \pm 0.5_5$	$0.29_5 \pm 0.07_5$	$0.22_3 \pm 0.05_0$	$0.21_2 \pm 0.03_7$	$0.24_4 \pm 0.6_4$
K_{270} (dL g^{-1} cm^{-1})	$0.16_0 \pm 0.01_5$	$0.13_8 \pm 0.04_4$	$0.13_2 \pm 0.03_7$	$0.16_0 \pm 0.01_0$	$0.16_4 \pm 0.01_0$	$0.14_8 \pm 0.03_2$

requisites for a probability (i.e., they should sum up to 1) (*40*). ANN consists of interconnected processing elements arranged in three kinds of layers named input, hidden, and output. Each processing element computes a weighted sum of the inputs from the previous layer (or from a data file for input layer neurons), transforms this sum in the output result by means of an activation function—usually a sigmoid—and propagates the result to the neurons in the successive layer or to the researcher, in the case of output neurons. All of the information is stored in the weights, which are the true memory of the system. In the back-propagation algorithm, each weight is adjusted iteratively, during the training phase, to minimize RMS error. This can be stated mathematically as

$$\Delta w_{ij}(t) = -\eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(t-1) \qquad (3)$$

where $\Delta w_{ij}(t)$ is the weight adjustment at the tth iteration, $\partial E/\partial w_{ij}$ is the partial derivative of the total RMS error with respect to the weight considered, and η and α are two coefficients, named learning rate and momentum, governing the rate of the training process and its stability (α can prevent the solution from being in local minima or damp the oscillations in the case of a large η) (*41*, *42*).

In this work, the statistical data processing was performed using Statistica for Windows v. 4.50 (StatSoft Inc., San Jose, CA) and NeuralWorks Professional II/Plus v. 5.30 (NeuralWare Inc., Pittsburgh, PA) packages.

RESULTS AND DISCUSSION

Variable Reduction. The main statistical variants pertinent to each olive cultivar and to the 153 oil samples as a whole (cumulative) are reported in Table 1. Peroxide value and some sterol percentages (Δ^5-avenasterol, brassicasterol, 24-methyl-enecholesterol, campestanol, Δ^7-campesterol, and $\Delta^{5,23}$-stigmastadienol) are absent because we excluded these indices on the basis of the following preliminary considerations:

(1) Peroxide value was available for only a too small number of samples, and we thought that filling the blanks with mean or extrapolated values, although possible, could have been misleading.

(2) Inspection of the correlation matrix proved the redundancy of Δ^5-avenasterol, which is linearly correlated to β-sitosterol ($\rho = 0.95$).

(3) The other rejected sterol percentages were all close to zero with CV > 100%.

From a first evaluation of the reported mean values and standard deviations of the chemical indices within each cultivar and the whole data set, it was not possible to formulate an immediate judgment on the discriminating ability of the single variable. To do this, and, if possible, to reduce further the number of chemical indices to be included in the classification/prediction model without compromising its discriminating ability, PCA and a variable selection method based on the Fisher F ratio were applied on our data set, which had been previously autoscaled (zero mean and unit variance) to exclude the variance ascribed to the different measurement units.

The Fisher F ratio provided the better results. In PCA, indeed, the first eight factors explained >80% of the total variance but did not allow an actual reduction as they were not identifiable with any of the original variables (chemical indices). Varimax rotation of the extracted factors provided this identification (factor loadings close to 1), but 13 factors were now necessary to explain the same percentage of the total variance (Table 2). On the contrary, 10 variables, corresponding to as many chemical indices, were selected by the Fisher F ratio method rejecting, at first, the variables with $F < 1$ and then, after an investigation upon their contribution to the classification/prediction model, all those with $F \leq 3.5$. As shown in Table 3, the 10 selected variables were acidity; palmitic (P), palmitoleic (P_o), stearic (S), oleic (O), linoleic (L), and linolenic (L_n) acids; sitostanol; Δ7-stigmastenol; and stigmasterol.

To perform a supervised analysis, this data set was split according to an "intelligent choice" into the training (115 samples) and the test (38 samples) sets, with a 3:1 ratio, taking care that samples from each cultivar and year of harvesting would be proportionally included in both. The 3:1 ratio was chosen to include in the model all of the possible sources of variation in a representative way. This division of samples has been performed three other times (115/38, 115/38, 114/39)—so that each sample would be represented in the test set only once—to evaluate prediction error by a cross-validation (bootstrap) procedure. In this procedure, the final prediction is made by comparing the results of the four prediction sets ($38 + 38 + 38 + 39 = 153$ predictions) with the actual 153 values.

An attempt to split the data set according to criteria different from those above-described—that is, selecting only the data from Customs' archives as training set and our data as test—has been rejected, as several samples were not correctly predicted by applying both LDA and ANN methods; particularly, all Frantoio samples were erroneously classified as Leccino.

Table 2. Principal Component Analysis: Factor Loading Matrix after Varimax Rotation

rotated factor	acidity	K_{270}	LLL	palmitic	palmit-oleic	stearic	oleic	lino-leic	lino-lenic	choles-terol	campes-terol	stigmas-terol	cleros-terol	β-sistos-terol	sitos-tanol	$\Delta^{5,24}$-stigmas-tadienol	Δ^{7}-sigmas-tenol	Δ^{7}-avenas-terol	eigen-value	% total variance
1	–0.20	–0.05	0.04	0.09	0.17	–0.12	–0.06	–0.19	–0.09	0.12	0.22	0.02	0.04	0.04	0.19	–0.09	**–0.90**	–0.53	1.36	**7.54**
2	–0.19	0.08	0.03	0.08	**0.79**	–0.12	0.24	–0.62	–0.16	0.10	0.23	–0.01	0.11	0.03	0.25	0.04	–0.14	–0.16	1.34	**15.01**
3	0.00	0.15	–0.03	0.02	–0.02	–0.03	–0.10	0.02	0.14	0.07	0.14	0.18	0.05	0.19	–0.03	**–0.91**	–0.09	–0.51	1.25	**21.94**
4	**0.93**	–0.03	0.04	0.03	–0.22	0.08	–0.05	0.30	0.06	–0.03	–0.26	0.07	0.02	0.00	–0.01	–0.02	0.21	0.16	1.17	**28.45**
5	–0.06	0.05	0.07	–0.05	0.17	0.01	0.27	–0.23	**–0.94**	0.02	0.02	–0.02	0.02	0.05	0.11	0.16	–0.10	–0.04	1.10	**34.56**
6	0.03	0.00	–0.11	0.02	–0.16	0.16	0.06	0.04	0.02	**–0.94**	–0.11	–0.05	–0.07	–0.16	–0.08	0.04	0.10	0.29	1.10	**40.66**
7	0.01	–0.07	–0.08	0.07	–0.09	0.03	0.11	–0.06	0.06	–0.17	0.03	–0.05	0.01	**–0.95**	0.01	0.19	0.01	0.28	1.10	**46.75**
8	–0.01	–0.16	0.06	0.08	0.27	–0.10	0.05	–0.29	–0.10	0.07	0.12	0.07	0.04	–0.01	**0.89**	0.04	–0.19	–0.05	1.08	**52.72**
9	0.03	**–0.94**	0.02	0.02	–0.11	–0.04	–0.24	0.06	0.05	0.00	–0.15	0.06	0.04	–0.06	0.20	0.16	–0.07	0.03	1.06	**58.63**
10	–0.02	0.04	–0.03	–0.10	–0.15	0.16	0.14	0.15	0.02	–0.07	–0.09	–0.03	**–0.96**	0.01	–0.05	0.05	0.05	0.04	1.05	**64.48**
11	0.05	–0.01	**0.96**	–0.08	0.04	–0.13	–0.06	–0.05	–0.08	0.12	0.22	0.02	0.03	0.08	0.07	0.04	–0.03	–0.06	1.05	**70.32**
12	0.09	–0.06	0.02	–0.07	–0.06	–0.02	–0.05	–0.06	0.03	0.05	0.11	**0.97**	0.04	0.06	0.10	–0.22	–0.03	–0.01	1.04	**76.10**
13	–0.07	–0.04	0.12	–0.01	0.14	**–0.93**	–0.06	–0.16	0.01	0.15	0.14	0.02	0.15	0.02	0.11	–0.03	–0.12	–0.08	1.04	**81.88**
14	–0.03	0.02	0.08	**–0.97**	–0.09	–0.01	–0.06	0.15	–0.06	0.02	0.02	0.07	–0.11	0.08	–0.10	0.03	0.12	–0.01	1.04	87.66
15	–0.04	0.18	–0.04	0.05	0.23	0.04	**0.86**	–0.28	–0.21	–0.04	0.04	–0.04	–0.10	–0.08	0.04	0.08	0.03	0.07	0.99	93.17
16	–0.17	0.10	0.13	–0.01	0.18	–0.10	0.02	–0.21	–0.01	0.08	**0.82**	0.07	0.06	–0.02	0.09	–0.07	–0.14	–0.21	0.90	98.20
17	–0.02	0.00	0.01	0.00	0.02	–0.01	–0.01	–0.03	–0.01	0.02	0.03	0.00	0.00	0.02	0.01	0.01	0.01	–0.41	0.18	99.17
18	–0.01	0.00	0.00	0.01	–0.10	–0.01	0.02	–0.37	–0.01	0.00	0.02	0.00	0.01	–0.01	0.02	0.01	–0.01	–0.02	0.15	100.00

Table 3. Fisher *F* Ratio Values Calculated for Each Chemical Index

variable	*F* ratio	variable	*F* ratio
stearic	**44.08**	K_{270}	3.36
palmitoleic	**38.80**	campesterol	2.43
linolenic	**31.88**	cholesterol	2.02
linoleic	**21.28**	Δ^{7}-avenasterol	1.84
oleic	**19.81**	β-sitosterol	1.27
palmitic	**9.70**	clerosterol	0.96
acidity	**6.23**	$\Delta^{5,24}$-stigmastadienol	0.96
Δ^{7}-stigmastenol	**5.25**	LLL	0.64
sitostanol	**5.09**		
stigmasterol	**3.79**		

Table 4. Linear Discriminant Analysis: Coefficients Assigned to Each Variable in the Five Classification Functions

variable	Carboncella	Frantoio	Leccino	Moraiolo	Pendolino
acidity	2.89	–1.45	–0.58	2.08	–3.52
palmitic	1.33	2.01	–1.47	–2.67	2.13
palmitoleic	–11.31	4.55	5.06	–6.73	7.95
stearic	0.12	–0.90	–2.84	–0.04	7.02
oleic	0.17	3.46	–3.79	1.24	–1.94
linoleic	4.48	1.27	–5.12	1.66	–2.48
linolenic	7.84	–2.05	–0.46	–0.20	–5.10
stigmasterol	–0.14	0.34	0.09	1.27	–1.82
sitostanol	–3.11	0.63	1.06	–0.34	1.05
Δ^{7}-stigmasterol	3.23	–0.46	–1.80	0.59	–1.71
constant	–18.61	–4.41	–8.03	–9.23	–17.35

Linear Discriminant Analysis. The mathematical model built by applying the LDA procedure on the 10 selected variables consists of five linear classification functions (c_i), one for each variety, of the form

$$c_i = c_{i1}v_1 + c_{i2}v_2 + ... + c_{in}v_n + c_{i0} \quad (4)$$

where v_1, ..., v_n are the values of each variable, c_{i1}, ..., c_{in} are the classification coefficients assigned to the respective variables, and c_{i0} is a constant (Table 4).

This model was able to identify all of the samples in the training set used to build itself and to correctly predict all of the individuals in the test set with a posterior probability of $\geq$0.95 for each of the four prediction sets; the only exceptions were a Frantoio oil in the first prediction set, having a posterior probability for its class of 0.55, and two Leccino oils in the third (posterior probabilities $\leq$ 0.75). An attempt to reduce further the number of chemical indices partly failed. In fact, the new models based on 9 (excluding acidity) or 7 (excluding sterols) variables were not able to correctly predict, respectively, 9 or 15 of the 153 cross-validation samples.

The predictive ability of this model can also be graphically visualized using linear discriminant functions, that is, a linear combination of the original variables, the optimal weights of which are obtained by maximizing the ratio of between-group variance of the individuals to their within-group variance. These functions can be thought of as the axes of a $(g - 1)$-dimensional orthogonal subspace—g being the number of classes—whereupon data points can be projected.

In our case all four extracted discriminant functions turned out to be significant, thus spanning a four-dimensional space. Figure 1 shows data point projection in the space of the first three discriminant functions (first sample division): it is evident that points are well grouped within single cultivars, which in turn are well separated.

Samples from different groups, seemingly superimposed in this figure (e.g., borderline samples from Carboncella and Moraiolo or from Frantoio and Leccino), are well separated in the plots obtained by considering the other three combinations of the four discriminant functions.

Artificial Neural Networks. To authenticate the varieties of our monocultivar olive oil samples, the ANN technique was applied, too. A feed-forward network with a single hidden layer and a back-propagation training algorithm (learning rule = generalized delta rule; transfer function = sigmoid; Softmax Output) was employed.

In a first attempt we used all of the variables selected for LDA analysis. A careful study and a series of tests were performed to optimize the network architecture and the values of the coefficients η (learning rate) and α (momentum). At last we opted for a 10−7−5 (number of nodes in input, hidden, and output layers respectively) network, with η = 0.450 (hidden → input) and 0.550 (output → hidden) and α = 0.400.

This network converged after ~3000 iterations with RMS = 0.0001, and it was able to exactly classify all of the samples in the training set and to correctly predict all those in the test set.

As the posterior probabilities assigned to each individual for its class were 1 for all of the analyzed samples, ANN analysis

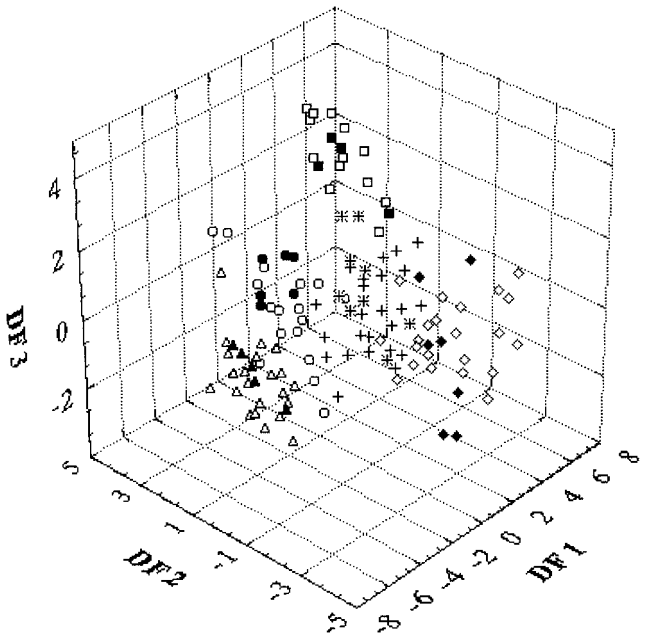

Figure 1. Linear discriminant analysis: mean scores of olive cultivars projected on the reduced space of the first three discriminant functions (first division of samples) [training and test set symbols: Carboncella (○, ●); Frantoio (+, ∗); Leccino (◇, ◆); Moraiolo (△, ▲); Pendolino (□, ■)].

was repeated using a reduced data set. Satisfactory results have been obtained retaining in the data set the six major fatty acids only (a further reduction of the variable number is seemingly superfluous because, the selected chemical indices being determinable on the same chromatogram, it would have allowed no saving to us of either time or expense).

The final 6−5−5 network, using the same values as above for η and α, gave a correct prediction for all of the test set samples, even if, in a few cases, posterior probabilities were <1.

Cross-validation on the other three divisions of samples resulted in a 100% predictive ability, but, also in these other cases, posterior probabilities for some samples were significantly <1.

Conclusions. In this work we have pointed out that it can be possible to discriminate extra virgin olive oils by their cultivar using only some of the chemical indices which can be determined according to the Official Analytical Methods provided for the commercial classification of this foodstuff. This should undoubtedly lead to advantages, both economic, as no additional cost for chemical analysis would be required, and organizational, as the results of the determination of these indices on a large number of samples, needed to build a representative data set, are available in the literature. The use of the conditional form is due to our restricting this analysis to only oil samples representative of the five cultivars from Sabina. The general validity of our conclusions should be checked by extending this investigation to oils extracted from different olive varieties or from the same varieties but obtained under different conditions (different region of origin, stage of ripeness, year of harvesting, or extracting procedure).

Limiting ourselves to the samples considered in the present work, we can affirm that 10 chemical indices are enough to lead to correct classification in LDA, whereas only six of them (major fatty acids, determined by a single chromatogram) are needed in ANN, which shows itself to be much more effective in solving this kind of problem.

However, attention should be paid to the selection of the training set, required to build the mathematical model (LDA) and to train the network (ANN). Indeed, the ability to correctly predict the class (variety) of a sample investigated to obtain the optimum calibration is highly affected by the representativeness of training set samples with respect to the whole data set. For this reason, the data eventually acquired from the literature must be integrated by the outcomes of the chemical analyses performed upon samples of authenticated origin and analogous to those under examination. The training set should be chosen from this data set on an "intelligent choice" basis (taking care that the different cultivars and the analyzed authentic origin samples are proportionally represented in both training and test sets). In fact, these conclusions stem from the unsatisfactory results obtained by splitting the data set according to criteria different from that stated above.

ACKNOWLEDGMENT

We acknowledge the oil producers from Sabina and particularly Dr. Gianfranco De Felici and Lara Fagiolo for having given us monocultivar oil samples.

LITERATURE CITED

(1) Zupan, J.; Novic, M.; Li, X.; Gasteiger, J. Classification of multicomponent analytical data of olive oils using different neural networks. *Anal. Chim. Acta* **1994**, *292*, 219−234.

(2) Lai, Y. W.; Kemsley, E. K.; Wilson, R. H. Potential of FTIR for the authentication of vegetable oils. *J. Agric. Food Chem.* **1994**, *42*, 1154−1159.

(3) Angerosa, F.; Di Giacinto, L.; Vito, R.; Cumitini, S. Sensory evaluation of virgin olive oils by ANN processing of Dynamic Headspace GC data. *J. Agric. Food Chem.* **1996**, *44*, 323−328.

(4) Bianchi, G.; Giansante, L.; Lazzari, M. Analisi per la tutela di genuinità, origine geografica e varietale degli oli vegetali. *Inf. Agric.* **1996**, *19*, 45−48.

(5) Shaw, A. D.; Di Camillo, A.; Vlahov, G.; Jones, A.; Bianchi, G.; Rowland, J.; Kell, D. B. 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* **1997**, *348*, 357−374.

(6) Bréas, O.; Guillou, C.; Reniero, F.; Sada, E.; Angerosa, F. ^{18}O measurements by continuous flow pyrolysis/Isotope ratio mass spectrometry of vegetable oils. *Rapid Commun. Mass. Spectrosc.* **1998**, *12*, 188−192.

(7) Vlahov, G.; Shaw, A. D.; Kell, D. B. Use of ^{13}C-NMR Distortionless Enhancement by Polarization Transfer pulse sequence and multivariate analysis to discriminate olive oil cultivars. *J. Am. Oil Chem. Soc.* **1999**, *76*, 1223−1231.

(8) Marini, D.; Balestrieri, F. Applicazione dell'analisi statistica multivariata alla differenziazione dell'origine degli oli di oliva. *Riv. Ital. Sci. Aliment.* **1994**, *23*, 361−366.

(9) Balestrieri, F.; Marini, D.; Salpietro, S. Analisi multivariata applicata alla differenziazione di oli di oliva extra vergine dell'Umbria. *Riv. Ital. Sci. Aliment.* **1995**, *24*, 15−22.

(10) Marini, D.; Balestrieri, F. Applicazione dell'analisi multivariata e delle reti neurali artificiali ai dati analitici degli oli di cartamo. *Riv. Ital. Sostanze Grasse* **1997**, *74*, 513−520.

(11) Balestrieri, F.; Marini, D.; Sacchini, A. Country-of-origin determination of cashew-nut. *Riv. Ital. Sostanze Grasse* **1993**, *70*, 11−20.

(12) Marini, D.; Balestrieri, F. Evaluation of wheat germ oils using artificial neural networks. *Riv. Ital. Sostanze Grasse* **1998**, *75*, 247−252.

(13) Forina, M.; Tiscornia, E. Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content. *Ann. Chim.* **1982**, *72*, 143−155.

(14) Leardi, R.; Paganuzzi, V. Characterization of the origin of extravirgin olive oils on the basis of sterol composition and statistical analysis. *Riv. Ital. Sostanze Grasse* **1987**, *64*, 131−136.
(15) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; Wiley: New York, 1986.
(16) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michette, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, The Netherlands, 1988.
(17) Brereton, R. G. *Chemometrics—Application of Mathematics and Statistics to Laboratory Systems*; Ellis Horwood: Chichester, U.K., 1990.
(18) Brereton, R. G. *Multivariate Pattern Recognition in Chemometrics*; Elsevier: Amsterdam, The Netherlands, 1992.
(19) EC (The Commission of the European Communities). Regulation 2568/91. *Off. J. Commission Eur. Communities* **1991**, *L248*, 1−83.
(20) Brown, P. J. *Measurement, Regression and Calibration*; Oxford Science Publications: Oxford, U.K., 1993.
(21) Sreerama, N.; Woody, R. W. Protein secondary structure from circular dichroism spectroscopy. Combining variable selection principle and cluster analysis with neural networks, ridge regression and self-consistent methods. *J. Mol. Biol.* **1994**, *242*, 497−507.
(22) Seasholtz, M. B.; Kowalski, B. R. The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* **1993**, *277*, 165−177.
(23) De Noord, O. E. Multivariate calibration standardization. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 65−70.
(24) Horst. *Factor Analysis of Data Matrices*; Hort, Rinehart and Winston: New York, 1965.
(25) Rummel, R. J. *Applied Factor Analysis*; Northwestern University Press: Evanston, IL, 1970.
(26) Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd ed.; Wiley: New York, 1991.
(27) Jollife, I. T. *Principal Component Analysis*; Springer: New York, 1986.
(28) Fukunaga, K. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press: San Diego, CA, 1990.
(29) Lee, P. M. *Bayesian Statistics: An Introduction*, 2nd ed.; Oxford University Press: New York, 1989.
(30) Fisher, R. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179−188.
(31) Cacoullos, T., Ed. *Discriminant Analysis and Application*; Academic Press: New York, 1973.
(32) McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley: New York, 1992.
(33) Coomans, D.; Massart, D. L.; Kaufman, L. Optimization by statistical Linear Discriminant Analysis in analytical chemistry. *Anal. Chim. Acta* **1979**, *112*, 97−122.
(34) Cherkassky, V., Friedman, J. H., Wechsler, H., Eds. *From Statistics to Neural Networks—Theory and Pattern Recognition Applications*; NATO ASI Series; Springer-Verlag: Berlin, Germany, 1994.
(35) Bishop, C. M. *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, U.K., 1995.
(36) Freeman, J. A.; Skapura, D. M. *Neural Networks: Algorithms, Applications and Programming Techniques*; Addison-Wesley: Reading, MA, 1991.
(37) Zupan, J.; Gasteiger, J. Neural Networks: a new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(38) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an Introduction*; VCH: Weinheim, Germany, 1993.
(39) Kröse, B.; Van Der Smagt, P. *An Introduction to Neural Networks*, 8th ed.; 1996; http://www.fwi.uva.nl/research/neuro.
(40) Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationship to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures and Applications*; Fougelman-Soulie, F., Hérault, J., Eds.; Springer-Verlag: Berlin, Germany, 1989; pp 227−236.
(41) Rumelheart, D. E.; Hinton, G. E.; Williams, R. J. Learning internal representation by error propagation. In *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*; Rumelheart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, MA, 1986; Vol. 1, pp 318−362.
(42) Rumelheart, D. E.; Durbin, R.; Golden, R.; Chauvin, Y. Backpropagation: the basic theory. In *Back-propagation: Theory, Architectures and Applications*; Chauvin, Y., Rumelheart, D. E., Eds.; Lawrence Erlbaum: Hillsdale, 1995; pp 1−34.

Received for review May 29, 2001. Revised manuscript received October 18, 2001. Accepted October 22, 2001. We acknowledge the National Research Council of Italy for financial support.

JF010696V