JUNE 2001
VOLUME 49, NUMBER 6

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Geographical Characterization of Italian Extra Virgin Olive Oils Using High-Field ^{1}H NMR Spectroscopy

Luisa Mannina,*,† Maurizio Patumi,‡ Noemi Proietti,§ Daniele Bassi,# and Anna Laura Segre§

Facoltà di Scienze, Università degli Studi del Molise, MM.FF.NN.,Via Mazzini 8, 86170 Isernia, Italy, and Istituto di Chimica Nucleare, CNR, 00016 Monterotondo Staz., Roma, Italy; Istituto Ricerche sull'Olivicoltura, CNR, Via Madonna Alta 128, 06100 Perugia, Italy; Istituto di Chimica Nucleare, CNR, 00016 Monterotondo Staz., Roma, Italy; and Dipartimento di Produzione Vegetale, Università di Milano, Sez. ne Coltivazioni Arboree, Via Celoria 2, 20133 Milano, Italy

^{1}H high-field nuclear magnetic resonance (NMR) was used to analyze 216 extra virgin olive oils collected in three years (1996, 1997, and 1998) in different Italian areas in order to evaluate the potential contribution of this technique to the geographical characterization of olive oils. A statistical procedure performed on the intensity of selected NMR peaks has been proposed. Tree clustering analysis of NMR data performed without any a priori hypothesis showed the existence of reliable parameters able to group the olive oils according to the location of olive oil production. Linear discriminant analysis applied to selected NMR parameters of olive oils of the same year of production allowed the grouping of samples according to their geographical origin with only very few errors. Moreover, a satisfactory grouping is reached by combining the NMR data of olive oils from two different years (1996 and 1997). Operating on appropriate sampling, a careful analysis of data yielded the conclusion that the place of olive production could be singled out as a discriminating factor regardless of the cultivars from which the olive oils are derived.

Keywords: *Proton NMR spectroscopy; geographical origin; cultivar; extra virgin olive oil; statistical analysis*

INTRODUCTION

The determination of the geographical origin of extra virgin olive oils is a rather recent problem: the quality of an olive oil is the result of different factors such as cultivar, environment, and cultural practices (*1*−*3*). Therefore, for the careful determination of the place of production based on chemical composition, many factors need to be taken into account (*4*).

Moreover, an important act of legislation, the "declared geographical origin" (*5*) allows the labeling of some European extra virgin olive oils with the names of the areas where they are produced. This certification improves the commercial value of the product. The supposed contribution of the area of olive production to the quality and the peculiarity of the olive oil is particularly important in Italy, where >200 different cultivars are grown in different areas.

Several attempts have been made to identify the place of olive oil production by means of multivariate analysis of suitable chemical parameters: using the principal component analysis (PCA) of fatty acid composition, Alessandri (*6*) and Forina and Tiscornia (*7*) obtained a first classification of Italian olive oils from different regions; Aparicio et al. (*8*−*11*), using an expert system

* Corresponding author (fax +39 0 6 906 72 477; e-mail nmr@mlib.cnr.it).
† Università degli Studi del Molise and Istituto di Chimica Nucleare, CNR.
‡ Istituto Ricerche sull'Olivicoltura, CNR.
§ Istituto di Chimica Nucleare CNR.
# Università di Milano.

10.1021/jf001408i CCC: $20.00 © 2001 American Chemical Society
Published on Web 05/31/2001

Figure 1. Map of Italy: the areas of production of the analyzed olive oils are shaded. 1996: Puglia, Sicily, and Liguria. 1997: Puglia, Sicily, and Liguria; 1998: Lazio, Lake Garda area, and Tuscany.

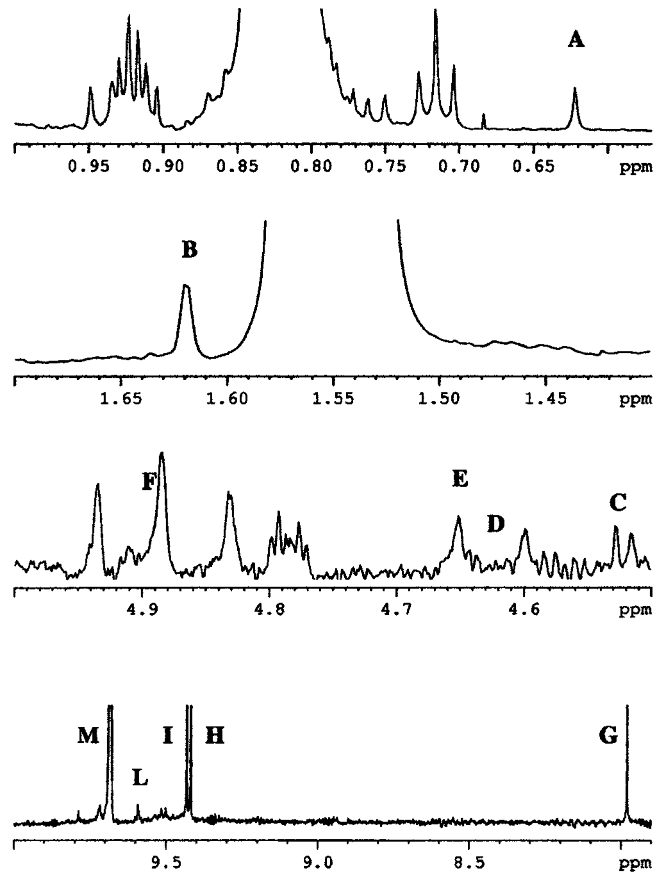

Figure 2. Expansion of several ^{1}H NMR spectral regions of an extra virgin olive oil. Labeled resonances, selected for the statistical analyses, are due to (A) β-sitosterol (0.622 ppm), (B) squalene (1.620 ppm), (C, D, E) terpernes (4.530, 4.627, and 4.654 ppm), (F) cycloartenol (4.886 ppm), (G) formaldehyde (8.007 ppm), (H) (*E*)-2-hexenal (9.450 ppm), (I, L) unsaturated aldehydes (9.540 and 9.610 ppm), and (M) hexanal (9.701 ppm).

(so-called SEXIA), have studied data from different chemical analyses to classify Spanish oils with respect to their origin and variety; Tsimidou et al. (*12*) applied the PCA of fatty acids and triacylglycerols for the geographical classification of Greek olive oils.

Recently, it has been shown that the combination of high-resolution NMR and statistical analysis gives interesting results for authentication purposes (*13*−*15*). In particular, a combined approach using NMR and gas chromatographic (GC) analyses has been proposed for the detection of the fraudulent additions of halzenut or sunflower oil to olive oil (*16*, *17*); the proposed methodology is based on the analysis of fatty acids in oils of different botanical origins.

Moreover, it has been shown that high-resolution ^{13}C NMR spectroscopy is able to provide valuable information about the acyl composition and the *sn* (strictly numbered)-1,3 and *sn*-2 acyl positional distribution of glycerol triesters in different vegetable oils (*18*−*21*).

In our previous papers it was shown that ^{1}H and ^{13}C high-field NMR spectroscopies give important results in the characterization and geographical classification of extra virgin olive oils (*20*−*26*). Preliminary studies were carried out on olive oils from different Italian regions, that is, Sicily, Campania, Lazio, and Umbria, and from different cultivars (*23*). Moreover, olive oils from different areas of the same Italian region, that is, Tuscany (*26*), and from other countries such as Spain and Argentina (research in progress) have been distinguished. In the case of olive oils from the same Italian region it was possible to discriminate between cultivar and environmental effect, showing that, after an accurate choice of the selection criteria, that is, selecting suitable resonances, the pedoclimatic effect was predominant (*26*).

In general, the proposed procedure seemed to be useful and promising. The method requires the ^{1}H NMR spectrum of the olive oil to be determined at high field and the intensity of a few selected normalized resonances to be measured. Of these resonances, the statistical processing of their relative weight allows a choice of the most significant ones. The 11 resonances chosen for the statistical analysis are due to the following minor components of olive oil: hexanal, *trans*-2-hexenal, two other unsaturated aldehydes, formaldehyde, three terpenes, squalene, cycloartenol, and β-sitosterol (*24*).

Only the resonances with the highest discriminant power, that is, with a good variability in many samples, must be taken in account. Because the variability of the 11 selected resonances is dependent on many different factors such as environment, cultivar, particular defects of the olive oil, and year or production, it is important to repeat this statistical analysis each time. This means that the correct resonances, that is, with the high discriminating power, must be identified according to the specific problem. For instance, a resonance due to a specific compound may be important in the discrimination of extra-European olive oils but not relevant within a European group. As an example of the above considerations, it was previously shown that in Tuscany olive oils squalene is the minor component with the major statistical weight for the geographical discrimination (*26*), whereas when olive oils from Argentina and Italy were compared, the amounts of β-sitosterol and linolenic acid (in preparation) are the most relevant parameters.

Table 1. Origins and Cultivars of Extra Virgin Olive Oils

origin	cultivar[a]	sample	origin	cultivar	sample	origin	cultivar	sample
				A. 1996				
Liguria	T	LI1	Sicily	Ce	SIC22	Puglia	O	PU9
Liguria	T	LI2	Sicily	B	SIC23	Puglia	O	PU10
Liguria	T	LI3	Sicily	B	SIC24	Puglia	O	PU11
Liguria	T	LI4	Sicily	Ce	SIC25	Puglia	O	PU12
Liguria	T	LI5	Sicily	B	SIC26	Puglia	O	PU13
Liguria	T	LI7	Sicily	B	SIC27	Puglia	O	PU14
Liguria	T	LI8	Sicily	Ce	SIC28	Puglia	O	PU15
Liguria	T	LI9	Sicily	N	SIC29	Puglia	O	PU16
Sicily	B	SIC1	Sicily	N	SIC30	Puglia	O	PU17
Sicily	B	SIC2	Sicily	Cr	SIC31	Puglia	O	PU18
Sicily	C	SIC3	Sicily	P	SIC32	Puglia	C, F, L, O	PU19
Sicily	B	SIC4	Sicily	C	SIC33	Puglia	C, L, CM	PU20
Sicily	B	SIC5	Sicily	C	SIC34	Puglia	C	PU21
Sicily	C	SIC6	Sicily	O	SIC35	Puglia	O+L	PU22
Sicily	N	SIC8	Sicily	B	SIC36	Puglia	C	PU23
Sicily	N	SIC9	Sicily	Mi	SIC37	Puglia	O	PU24
Sicily	Cr	SIC10	Sicily	Mi	SIC38	Puglia	C, L, CM	PU25
Sicily	P	SIC11	Sicily	V	SIC39	Puglia	O, L	PU26
Sicily	P	SIC12	Puglia	O	PU1	Puglia	C, L, O, F	PU27
Sicily	Ce	SIC13	Puglia	O	PU2	Puglia	C	PU28
Sicily	Ce	SIC14	Puglia	O	PU3	Puglia	O	PU29
Sicily	B	SIC16	Puglia	O	PU4	Puglia	O, L	PU30
Sicily	Mi	SIC18	Puglia	O	PU5	Puglia	C, O, L, F	PU31
Sicily	V	SIC19	Puglia	O	PU6	Puglia	C, L, CM	PU32
Sicily	Ce	SIC20	Puglia	O	PU7			
Sicily	Ce	SIC21	Puglia	O	PU8			
				B. 1997				
Liguria	T	LI65	Sicily	No	SIC 9	Puglia	C	PU58
Liguria	F, L	LI66	Sicily	No	SIC 10	Puglia	O	PU59
Liguria	T	LI67	Sicily	No	SIC 11	Puglia	C, L, CM	PU60
Liguria	T	LI76	Sicily	No	SIC 12	Puglia	C, L, O, N	PU61
Liguria	F	LI77	Sicily	No	SIC 13	Puglia	O, L	PU62
Liguria	T	LI78	Sicily	No	SIC 14	Puglia	C	PU69
Liguria	T	LI87	Sicily	No	SIC 15	Puglia	C	PU70
Liguria	T	LI88	Sicily	No	SIC 23	Puglia	O, L	PU71
Sicily	No	SIC1	Sicily	No		Puglia	O	PU73
Sicily	No	SIC2	Sicily	No		Puglia	C, L, CM	PU74
Sicily	No	SIC3	Sicily	No		Puglia	C, L, CM	PU81
Sicily	No	SIC4	Sicily	No		Puglia	O, L	PU82
Sicily	No	SIC5	Sicily	No		Puglia	O, L	PU83
Sicily	No	SIC6	Sicily	No		Puglia	C	PU84
Sicily	No	SIC7	Sicily	T I		Puglia	O	PU85
Sicily	No	SIC8						
				C. 1998				
Arezzo, Tuscany	F, L, M, Ne	TUAR1	Seggiano, Tuscany	S	TUS4	Lake Garda	Rossanello	GAR 39
Arezzo, Tuscany	F, L, M, Ne	TUAR2	Seggiano, Tuscany	S	TUS5	Lake Garda	Tr	GAR 40
Arezzo, Tuscany	F, L, M, Ne	TUAR3	Seggiano, Tuscany	S	TUS6	Lazio	Si, R	LA1
Arezzo, Tuscany	F, L, M	TUAR4	Seggiano, Tuscany	S	TUS7	Lazio	Si, F, L, Ma	LA2
Arezzo, Tuscany	F, L, M	TUAR5	Seggiano, Tuscany	S	TUS8	Lazio	Si, F, L, Ma	LA3
Arezzo, Tuscany	F, L, M	TUAR6	Seggiano, Tuscany	S	TUS9	Lazio	F, L	LA4
Arezzo, Tuscany	F +M+ P	TUAR7	Seggiano, Tuscany	S	TUS10	Lazio	Si, R	LA5
Arezzo, Tuscany	F, L, M, Ne	TUAR8	Seggiano, Tuscany	S	TUS11	Lazio	Si, R	LA6
Arezzo, Tuscany	F, L, M	TUAR9	Seggiano, Tuscany	S	TUS12	Lazio	Si, R	LA7
Arezzo, Tuscany	F, M, L	TUAR10	Seggiano, Tuscany	S	TUS13	Lazio	Si, R, F, L	LA8
Arezzo, Tuscany	F, M, L	TUAR11	Seggiano, Tuscany	S	TUS14	Lazio	Si	LA9
Arezzo, Tuscany	F, M, L	TUAR12	Seggiano Tuscany	S	TUS15	Lazio	Si, F, L	LA10
Arezzo, Tuscany	F, M, L	TUAR13	Lake Garda	L	GAR 21	Lazio	F, L, I77, I79, Si	LA11
Lucca, Tuscany	F, M	TUA1	Lake Garda	P	GAR 22	Lazio	F, L	LA12
Lucca, Tuscany	F, L	TUA2	Lake Garda	Ba	GAR 23	Lazio	F, L	LA13
Lucca, Tuscany	F	TUA3	Lake Garda	Ca2	GAR 24	Lazio	F, L	LA14
Lucca, Tuscany	F, L	TUA4	Lake Garda	Ca1	GAR 25	Lazio	F, L	LA15
Lucca, Tuscany	Q	TUA5	Lake Garda	Co	GAR 26	Lazio	F, L	LA16
Lucca, Tuscany	Q	TUA6	Lake Garda	Fa2	GAR 27	Lazio	F, L	LA17
Lucca, Tuscany	F	TUA7	Lake Garda	F	GAR 28	Lazio	F, L	LA18
Lucca, Tuscany	L	TUA8	Lake Garda	Ga	GAR 29	Lazio	F, L	LA19
Lucca, Tuscany	F	TUA9	Lake Garda	Gr	GAR 30	Lazio	F, L	LA20
Lucca, Tuscany	Q	TUA10	Lake Garda	L	GAR 31	Lazio	F, L	LA21
Lucca, Tuscany	F, L	TUA11	Lake Garda	Le	GAR 32	Lazio	F, L	LA22
Lucca, Tuscany	Q	TUA13	Lake Garda	Min2	GAR 33	Lazio	F, L	LA23
Lucca, Tuscany	Q	TUA14	Lake Garda	Mt	GAR 34	Lazio	F, L	LA24
Lucca, Tuscany	F	TUA15	Lake Garda	Ma	GAR 35	Lazio	Si	LA25
Seggiano, Tuscany	S	TUS1	Lake Garda	Pe	GAR 36	Lazio	Si	LA26
Seggiano, Tuscany	S	TUS2	Lake Garda	Ra	GAR 37	Lazio	F, L	LA27
Seggiano, Tuscany	S	TUS3	Lake Garda	Re	GAR 38	Lazio	F, L	LA28

[a] B, Biancolilla; C, Coratina; Ce, Cerasuola; CM, Cima di Mola; Cr, Crastu; F, Frantoio; L, Leccino; M, Moraiolo; Mi, Minuta; N, Nocellara del Belice; O, Ogliarola; P, Passalunara; T, Taggiasca; V, Verdello; No, Nociara; TI, Tonda Iblea; Ba, Baia; Ca1, Casaliva 1; Ca2, Casaliva 2; Co, Cornarol; Fa2, Favarol 2; Ga, Gargnà; Gr, Grignano; Le, Less; Ma = Maurino; Mt = Mitria; Min2, Miniol 2; Ne, Nerino; Q, Quercetana; Ra, Raza; R, Reale; Re, Regina; S, Seggianese; Si, Sirole; P, Pendolino; Tr, Trep.

Table 2. ANOVA of the Selected Intensity NMR Data[a]

selected resonances (ppm)	1996–1997 olive oils from PU, SIC, and LI		1997–1998 olive oils from PU, SIC, and LI		1998–1999 olive oils from LA, TU, and GAR	
	F (2.73)	*p* level	*F* (2.73)	*p* level	*F* (2.73)	*p* level
0.620	5.67	0.005	19.39	0.000001	14.35	0.000001
1.620	41.19	0.000001	40.20	0.000001	90.46	0.000001
4.530	8.015	0.000713	6.261	0.004107	16.62	0.000001
4.627	20.89	0.000001	7.284	0.001887	11.53	0.000001
4.654	3.827	0.026256	16.91	0.000004	66.56	0.000001
4.886	18.12	0.000001	10.04	0.000264	135.7	0.001
8.007	5.139	0.008162	4.731	0.0135887	41.09	0.000001
9.450	15.18	0.000003	2.914	0.064616	216.7	0.001
9.540	0.2875	0.750958	7.902	0.001195	113.9	0.001
9.610	4.594	0.013195	6.227	0.004220	525.5	0.001
9.701	13.20	0.000013	4.091	0.023653	377.3	0.001

[a] See Appendix.

In this paper, we report the statistical criteria to be used to obtain a clear geographical separation among Italian extra virgin olive oils.

MATERIALS AND METHODS

Sampling. The origins and cultivars of 216 extra virgin olive oil samples collected in three years (1996, 1997, and 1998) and produced in different areas of Italy (see Figure 1) are reported in Table 1. The sampling of 1996 and 1997 consists of monovarietal and multivarietal olive oils from Sicily, Puglia, and Liguria. The sampling of 1998 consists of olive oils from nearby geographical regions, that is, Tuscany and Lazio, olive oils from cultivars grown in a particular environment, that is, Lake Garda, and olive oils from a local cultivar, Seggianese, grown in a borderline district between Lazio and Tuscany.

NMR Analysis. The NMR procedure previously reported by Segre and Mannina (*24*) was followed. Olive oils (20 μL) were placed into 5 mm NMR tubes and dissolved in chloroform-*d* (0.7 mL) and DMSO-*d* (20 μL). ^{1}H NMR spectra were recorded on a Bruker AMX 600 (Karlsruhe, Germany) instrument operating at 600.13 MHz. The deuterium signal of $CDCl_3$ was used to lock the magnetic field.

^{1}H NMR spectra were obtained using the following acquisition parameters: $\pi/2$ pulse; acquired points, 32K; processed points, 32K; spectral width, 14 ppm; acquisition time, 1.5 s; relaxation delay, 2 s; number of scans, 4000. Before the quantitative evaluation of all peaks of interest was performed, a careful baseline correction was performed. The intensities of the selected resonances (see Figure 2) were compared to that of the methylene resonance at 1.553 ppm, the intensityof which is set to 1000. This normalizing procedure gives for each resonance an index proportional to the molar ratio between each compound and the total amount of the fatty chains.

Statistical Methods. NMR data were submitted to Statistica software package for Windows (1997 edition by Statsoft, Inc.).

A statistical procedure based on five points was followed (see also the Appendix):

1. *Analysis of variance (ANOVA)* on the selected resonances;

2. *Tree clustering analysis (TCA)* on 11 selected resonances without any a priori hypothesis;

3. *K-means clustering analysis (LDA)* on the 11 selected resonances with the a priori hypothesis, that is, the number of places of production on the basis of the TCA;

4. *Linear discriminant analysis (LDA)* on the 11 resonances with the a priori hypothesis, that is, the number of places of production on the basis of the TCA (the highest discriminating power of different methods is not necessarily obtained with the same resonances); and

5. *Reliability of the system*. To prove the reliability of the system, some randomly selected olive oil samples are not included in the statistical analysis and are considered as unknown samples in further calculation. If the selected unknown samples are well classified, the system is stable and can be used for real samples.

RESULTS AND DISCUSSION

In Figure 2, several expansions of the ^{1}H NMR spectrum of an extra virgin olive oil are reported; the 11 selected resonances used for the statistical analyses are labeled.

Statistical results corresponding to each year will be discussed separately.

1996. Seventy-six extra virgin olive oils from Liguria, Sicily, and Puglia have been analyzed: the intensities of 11 selected resonances have been submitted to the following statistical analyses.

ANOVA. This analysis was applied to the 11 selected NMR resonances. The results, summarized in Table 2, show that 10 of the selected variables were significantly different for each region; only 1 variable at 9.54 ppm is not relevant for discrimination purposes because it does not have a significant variability in the three regions.

TCA. The results of this analysis are reported as a dendrogram in Figure 3. This dendrogram shows a grouping according to the geographical areas. Cutting the tree at an appropriate level, three groups are obtained: two groups consist of 8 and 22 olive oils from Sicily; the third group consists of olive oils from Liguria and Puglia. Cutting the tree at a further level, the seven samples from Liguria are grouped together. All 76 samples are correctly classified with the exception of 6 samples from Sicily (SIC13, SIC12, SIC20, SIC23, SIC36, and SIC38) and 2 from Puglia (PU31 and PU32). Note that, even if these samples are misplaced, some of them are in a border position not distant from the correct group. Altogether the classification error is <10%.

K-Means Clustering. The obtained results of this statistical approach show that three different groups have been obtained: cluster 1, with samples from only Sicily; cluster 2, with samples mostly from Sicily; and cluster 3, with samples from Puglia and Liguria.

LDA. Olive oils from the same region are well grouped: the ellipses delimit 95% confidence (see Figure 3). It is possible to observe that only three olive oils from Sicily are not correctly classified. The discriminating power of the selected variables is given by Wilks' lambda factor (see Table 3 and Appendix). This parameter is near zero for the variables with a high discriminant power. In this case all of the selected resonances have a similar discriminant powers.

To prove the reliability of the model, the method has been checked using known samples as unknown vari-

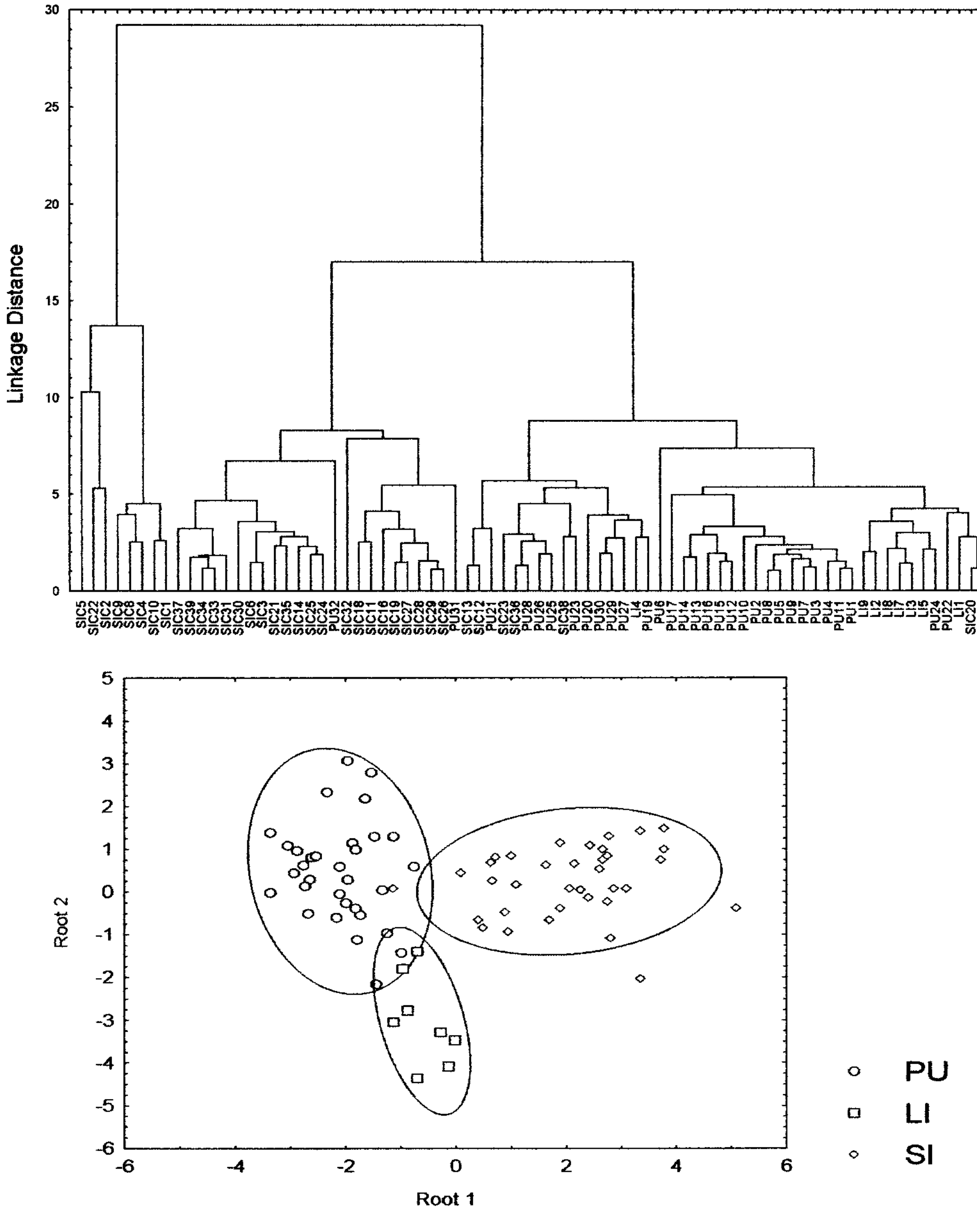

Figure 3. TCA (dendrogram) and LDA of extra virgin olive oils from three Italian regions in 1996. For TCA samples labeled with the same letter are from the same region: LI, Liguria; PU, Puglia; SIC, Sicily. For LDA canonical scores for the two discriminant equations (roots 1 and 2) are reported. Ellipses represent the 95% confidence regions for each group. Samples labeled with the same symbol are from the same region: ○, Puglia; □, Liguria; ◇, Sicily.

Table 3. LDA of ^{1}H NMR Data from Extra Virgin Olive Oils from Three Italian Regions (Liguria, Puglia, and Sicily) in 1996

LDA variable (ppm)	raw coefficients for canonical variables		Wilks' lambda factor for the model without the selected variable
	root 1	root 2	
0.622	0.60669	0.045268	0.110443
1.620	0.67121	0.265177	0.118867
4.530	0.490709	−0.602354	0.117437
4.627	−0.583522	0.839815	0.126932
4.654	0.146279	−0.571473	0.106881
4.886	0.857106	−0.077031	0.147286
8.007	0.442809	−0.256488	0.099506
9.450	−0.179397	0.246779	0.094761
9.610	0.154490	0.209776	0.093832
9.701	−0.538952	−0.644369	0.106730
eigenvalue	4.048750	1.146996	
cum prop	0.779243	1.000000	

ables. In detail, three times different and randomly selected sets of olive oils composed of five olive oils from Puglia and five from Sicily were removed from the data and the model was calculated again. The three Sicily samples not properly classified were not included. No olive oils from Liguria have been chosen to test the model because this group is constituted by a small number of samples.

The excluded olive oils were introduced to the system as unknowns. In all runs, all samples were correctly classified, so the system is stable and can be used for real samples.

1997. Forty-six extra virgin olive oils from Liguria, Sicily, and Puglia have been analyzed: the intensities of 11 selected resonances have been submitted to the following statistical analyses.

ANOVA. The ANOVA was applied to the 11 selected resonances of the NMR data. The results reported in Table 2 show that all 11 selected variables were significantly different for the three regions.

TCA. The results are reported in Figure 4. On the left

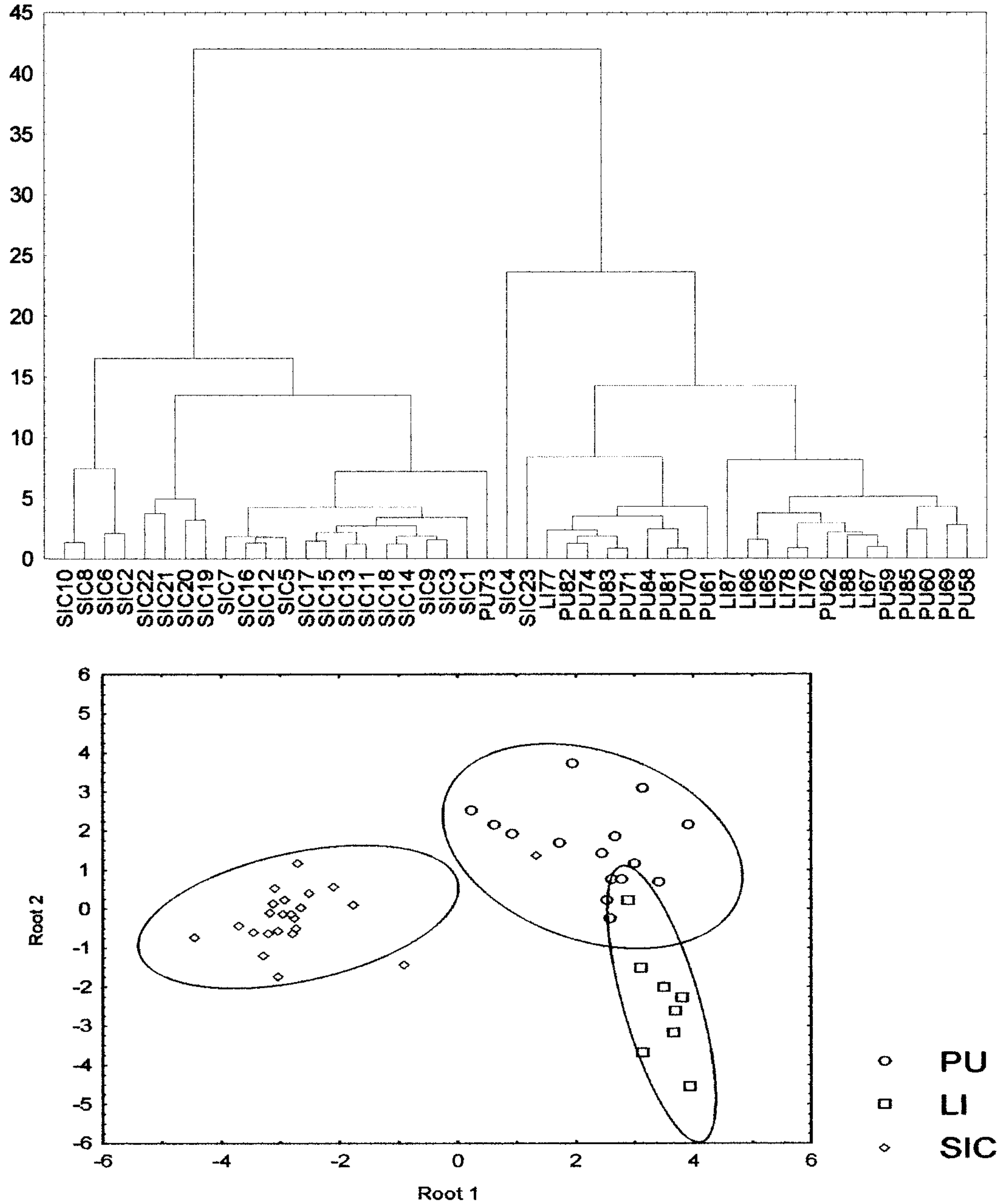

Figure 4. TCA (dendrogram) and LDA of extra virgin olive oils from three Italian regions in 1997 (Liguria, Puglia, and Sicilia). For TCA samples labeled with the same letter are from the same region (LI, Liguria; PU, Puglia; SIC, Sicily). For LDA canonical scores for the two discriminant equations (roots 1 and 2) are reported. Ellipses represent the 95% confidence regions for each group. Samples labeled with the same symbol are from the same region: ○, Puglia; □, Liguria; ◇, Sicily.

a large branch with 21 Sicilian olive oils is observed. A contiguous group contains mostly olive oils from Puglia and Liguria. A further cut allows the discrimination of only eight oils from Puglia with a few misplaced samples (SIC4 and SIC 23 on the borderline and LI77). Within this group, six olive oils from Liguria are grouped together, with the only exception the misplaced sample PU62. The total geographical errors is ∼15%.

These results suggest that the number of geographical sites where olive oils have been produced can be obtained without any a priori hypothesis and subsequently can be used as an *input* for all other statistical methods.

K-Means Clustering. The results of this statistical analysis are reported as analysis of between- and within-group variance and Euclidean distance between clusters (see Appendix).

Three clusters were obtained: cluster 1, which groups olive oils mainly from Puglia; cluster 2, which is composed of olive oils from Sicily; and cluster 3, which includes olive oils from Puglia and Liguria.

LDA. Olive oils from the same regions are well grouped, with the shown ellipses delimiting the 95% confidence (Figure 4**)**. Only two olive oils from Sicily are not correctly classified. The discriminating power of selected variables is given by Wilks' lambda factor (Table 4). The reliability of the system has been proven by using the procedure reported for 1996. In detail, five olive oils from Puglia and five from Sicily were removed as unknown samples. The obtained results show that the system is stable and can be used for real samples.

1996—1997. An interesting question is if it is possible to create a data bank for the geographical discrimination of olive oils and if it is necessary to regenerate this data

Table 4. LDA of ^{1}H NMR Data from Extra Virgin Olive Oils from Three Italian Regions (Liguria, Puglia, and Sicily) in 1997

LDA variable (ppm)	raw coefficients for canonical variables		Wilks' lambda factor for the model without the selected variable
	root 1	root 2	
0.622	0.918117	1.02372	0.0711767
1.620	−0.796623	−0.12571	0.061014
4.530	0.298545	−0.60718	0.041624
4.627	−0.768803	−0.25091	0.039916
4.654	−0.354408	−1.06653	0.060420
4.886	−0.406661	0.38843	0.040539
8.007	−0.103027	0.57597	0.039172
9.450	−0.029251	−0.30164	0.037272
9.540	−0.228305	−0.59209	0.038046
9.610	−0.267437	1.35802	0.041266
9.701	0.765614	−0.93267	0.043105
eigenvalue	7.929046	2.02120	
cum prop	0.796870	1.000000	

bank every year or if common criteria exist that allow the extrapolation of data in different years.

To answer these questions, statistical methods were applied to both years 1996 and 1997.

The results of the LDA are reported in Figure 5. It is important to observe that although the total number of errors is larger than the analysis applied within a single year, still an adequate geographical classification is obtainable.

1998. Ninety extra virgin olive oils from three different areas of Tuscany, the Lake Garda area, and Lazio have been analyzed: the intensities of 11 selected resonances have been submitted to the following statistical analyses.

ANOVA. The ANOVA was applied to the 11 selected intensities of the NMR data (Table 2). The results show that all 11 selected variables were highly significant for discrimination purpose.

TCA. The results are reported in Figure 6. When the cluster is cut at an appropriate level, the following groups corresponding to different geographical areas of production are obtained: a group with samples from the Arezzo district **(**Tuscany**)**, from Lazio, from the Lucca district **(**Tuscany**)**, from Lake Garda, and from the Seggiano district (Tuscany). It is to be noted that olive oils from Arezzo, a district in southern Tuscany, are linked in the same big cluster together with samples from the nearby Lazio, although distinctly grouped by place of production despite the cultivars being mainly the same (Frantoio and Leccino). The geographical separation among the above olive oils and those from the Seggiano or Lucca districts **(**Tuscany**)** is always excellent (*26*), although in this case the cultivar and environment effects cannot be easily separated.

All samples are correctly classified with the exception of GAR33, GAR30, and TUAR2, although they are grouped in border positions.

LDA. Olive oils from the same place of production were well grouped: the ellipses delimit 95% confidence (Figure 6). The discriminating power of selected variables is given by Wilks' lambda factor (Table 5). The reliability of the system has been previously reported.

In detail, three different sets of oils composed to two olive oils from Tuscany (Arezzo district), two olive oils from Tuscany (Seggiano district), two olive oils from Tuscany (Lucca district), eight olive oils from Lazio, and three from Garda were removed. The excluded oils were then introduced to the statistical analysis as unknowns, and all samples were classified correctly; thus, the system is stable and can be used for real samples.

The statistical analysis distinguished the pedoclimatic factor from the cultivar effect. Particularly striking was the sound differentiation between samples of similar genetic origins from two close places of production (Arezzo and Lazio) and the strong separation of the oils from Lake Garda from oils from all other regions. Furthermore, despite the rather large genetic diversity present in Garda olive oils, all samples were very tightly grouped by the chosen ^{1}H NMR parameters.

The investigation carried out in the present work offers a reliable protocol that can be extended to olive oil characterization when the geographical origin must be identified. The present findings suggest a positive contribution of ^{1}H NMR analysis to the characterization of extra virgin olive oils and certification of the geographical origin (D.O.P.), regardless the cultivar. Moreover, the above data can be used to create a data bank that would be usable for verifying the geographical origin of olive oils.

APPENDIX

Statistical Analysis (*27, 28*). The normalized intensities of the selected ^{1}H resonances have been submitted to ANOVA: it proves that the null hypothesis (i.e., no statistically significant differences between the variances of the groups) for the selected resonances is not valid. The results of this analysis are reported as F value and p level. The F value with the degrees of freedom tests whether the between and within variances are significantly different. The p level represents a decreasing index of the reliability of a result and gives the probability of error involved in accepting a result as valid. A p level of ≤ 0.05 (5% probability of error) is usually treated as a borderline acceptable error level.

Different methods of classification have been applied: *TCA*, *K-means clustering*, and *LDA* (*28*). These methods rely on different basic ideas. TCA, unlike many other statistical procedures, is mostly used when no a priori hypothesis is given and the procedure finds the most significant possible solution: the main purpose is to cluster the data into meaningful groups. In the present work, it was important to start the analysis without any a priori hypothesis to verify whether the observed classification reflected the cultivar or the environmental effect. The results of this analysis can be used as a priori data for further analyses such as the K-mean and the LDA.

TCA. The tree clustering method joins together objects (olive oils) into successively larger clusters, using some measure of distance.

Usable distances are Euclidean, square Euclidean, Chebychev, and Power. The Euclidean distance is the geometric distance in the multidimensional space; it is computed as

$$\text{distance}\ (x, y) = [\Sigma_i (x_i - y_i)^2]^{1/2}$$

The squared Euclidean distance is used to place progressively greater weight on objects that are apart; this distance is computed as

$$\text{distance}\ (x, y) = \Sigma_i (x_i - y_i)^2$$

The Chebychev distance is used to define objects as "different" if they are different in any one of the

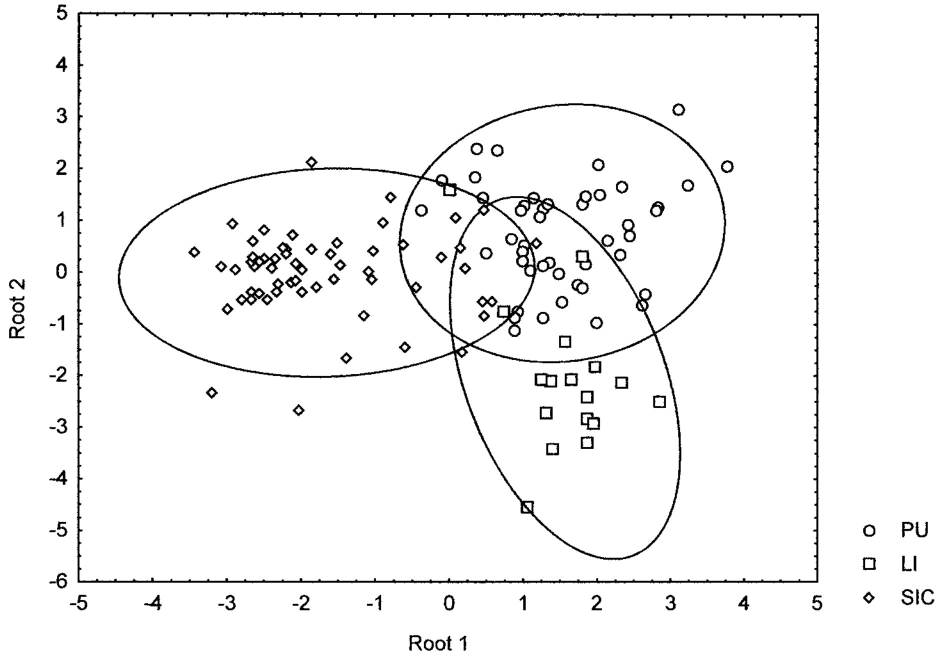

Figure 5. LDA of extra virgin olive oils of 1996 and 1997. Canonical scores for two discriminant equations (roots 1 and 2) are reported. Ellipses represent the 95% confidence regions for each group. Samples labeled with the same symbol come from the same region: ○, Puglia; ◇, Sicily; □, Liguria.

Table 5. LDA of [1]H NMR Data from Extra Virgin Olive Oils from Five Italian Areas (Arezzo, Lucca, and Seggiano Districts in Tuscany, Lazio Region, and Lake Garda) in 1998

LDA variable (ppm)	raw coefficients for canonical variables		Wilks' lambda factor for the model without the selected variable
	root 1	root 2	
0.622	0.04129	0.31799	0.000061
1.620	−0.30097	−0.20612	0.000095
4.530	0.14485	0.05242	0.000059
4.627	0.18566	0.45404	0.000074
4.654	0.01752	−0.69258	0.000135
4.886	0.47737	0.55937	0.000167
8.007	−0.29721	0.27820	0.000075
9.450	−0.05045	−1.05851	0.000151
9.540	0.02365	0.27834	0.000065
9.610	−0.81481	−0.03963	0.000110
9.701	0.78768	0.33623	0.000079
eigenvalue	65.17258	4.06556	
cum prop	0.73788	0.97745	

dimensions; this distance is computed as

$$\text{distance}\ (x, y) = \text{maximum}\ |x_i - y_i|$$

The Power distance is used to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different; this distance is computed as

$$\text{distance}\ (x, y) = [\Sigma_i |x_i - y_i|^p]^{1/r}$$

where r and p are user-defined parameters: parameter p controls the progressive weight that is placed on differences on individual dimensions; parameter r controls the progressive weight that is placed on larger differences between objects. If r and p are equal to 2, then this distance is equal to Euclidean distance. In our analysis $p = 2$ and $r = 1$, 2, or 3.

Given n olive oils characterized by p parameters, the problem is to classify the olive oils into homogeneous groups (clusters). When the distance function is changed, different classifications can be obtained. The statistical results are reported as a dendrogram having leaves that are the n olive oils. All possible classifications in a prescribed number of groups can be obtained by cutting the tree at a suitable level. All of the above distances in our case give rather similar results, confirming the conservative nature of the data.

In the tree clustering method, a linkage or amalgamation rule is necessary to determine when two clusters are sufficiently similar to be linked together. There are different linkage rules such as single linkage, complete linkage, and unweighted pair-group average.

In the single-linkage method the distance between two clusters is calculated by the distance of the two closest objects in the different clusters.

In the complete linkage method the distance between clusters is determined by the greatest distance between any two objects in the different clusters.

In the unweighted pair-group average the distance is calculated as the average distance between all pairs of objects in the two different clusters. This method is very efficient when objects form naturally distinct clumps and with "chain" type clusters. Using all of these methods, we obtained the same type of clustering, data not shown.

The *K-means clustering* is applied when a hypothesis concerning the number of clusters in the variables is given; it produces K different clusters with the greatest possible distinction. This procedure moves objects around from cluster to cluster with the purpose of minimizing the within-cluster variance and maximizing the between-cluster variance. The K-means clustering results are reported as analysis of between- and within-group variance and Euclidean distance between clusters.

In the ANOVA, the between-group variance is compared to the within-group variance to decide whether the means for a particular variable are significantly different between groups.

The LDA is a multivariate method used to determine which variables discriminate better between two or

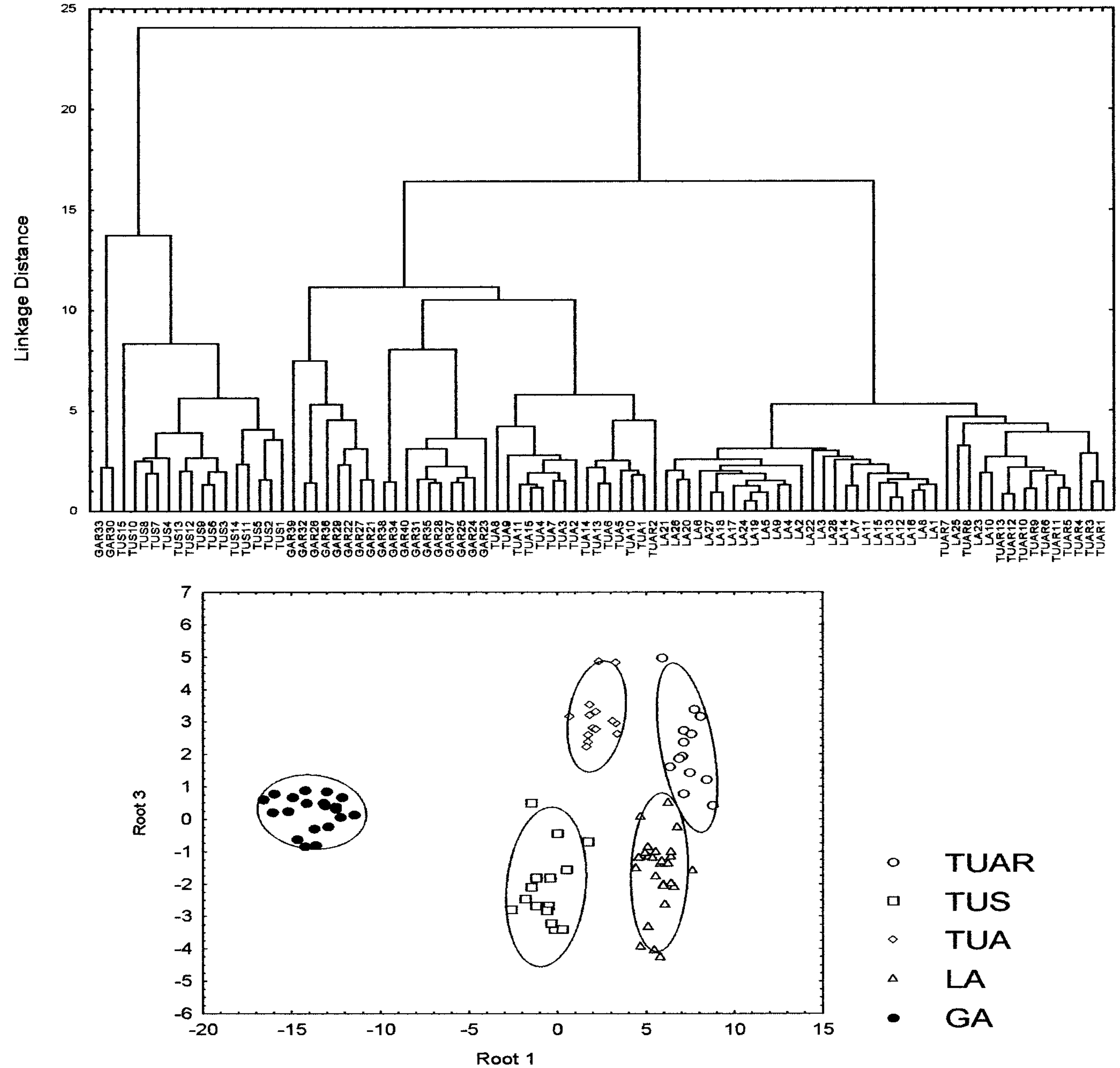

Figure 6. TCA (dendrogram) and LDA of extra virgin olive oils from different Italian areas in 1998 (Lake Garda, Lucca, Arezzo, Seggiano, and Lazio). For TCA samples labeled with the same letter are from the same region (GAR, Garda; TUAR, Arezzo; TUA, Lucca; TUS, Seggianese; LA, Lazio). For LDA canonical scores for the two discriminant equations (roots 1 and 2) are reported. Ellipses represent the 95% confidence regions for each group. Samples labeled with the same symbol are from the same region: ●, Lake Garda; ○, Arezzo, Tuscany; ◇, Lucca, Tuscany; □, Seggiano, Tuscany; △, Lazio.

more a priori defined groups**.** This procedure needs selected variables to build up a data matrix and to give rise to discriminant (canonical) linear functions. The number of functions is equal to the number of groups minus one; in the present work two discriminant functions have been estimated: canonical variables and Wilks' lambda factors. The latter parameter gives the discriminating power of selected variables, and its value ranges from 1.0 (no discriminatory power) to 0.0 (perfect discriminatory power); the value after the selected variable has been removed is reported.

To obtain stable systems for the unknown samples, it is important to have many test samples for each group; in fact, LDA applied to a few samples can give an artificially good separation, and the unknown groups can be wrongly classified. To prove the reliability of the model, the system has been checked using known samples as unknown samples. Three times a different and randomly selected set of oils for each year has been removed from the data. With the remaining data the model has been calculated again. The excluded oils were then introduced to the statistical analysis as unknowns; if the unknown samples are correctly classified, the system is stable and can be used for real samples.

LITERATURE CITED

(1) Solinas, M.; Marsilio, V.; Angerosa, F. Behaviour of some components of virgin olive oil flavour in relation to olive ripening. *Riv. Ital. Sostanze Grasse* **1987**, *64*, 475−480.
(2) Solinas, M. HRGC analysis of virgin olive oil phenolic compounds in relation to olive ripening degree and variety. *Riv. Ital. Sostanze Grasse* **1987**, *64*, 255−262.

(3) Fontanazza, G.; Patumi, M.; Solinas, M.; Serraiocco, A. Influence of cultivars of the composition and quality of olive oil. *Acta Hortic.* **1993**, *356*, 358−361.
(4) Montedoro, G. F.; Servili, M. I parametri di qualità dell'olio di oliva ed i fattori agronomici e tecnologici che li condizionano. *Riv. Ital. Sostanze Grasse* **1992**, *69*, 563−568.
(5) European Comunity. Regulation 2081/92. *Off. J. Eur. Communities* **1992**.
(6) Alessandri, S.; Cimato, A.; Modi, G.; Mattei, A.; Crescenzi, A.; Caselli S.; Tracchi, S. Univariate models to classify Tuscan virgin olive oils by zone. *Riv. Ital. Sostanze Grasse* **1997**, *74*, 155−163.
(7) 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.
(8) Aparicio, R.; Albi, T.; Lanzon, A.; Navas, M. A. SEXIA: an expert system to oils identification database from olive grove zones. *Grasas Aceites* **1987**, *38*, 9−14.
(9) Aparicio, R.; Albi, T.; Cert, A.; Lanzon, A. SEXIA expert system: canonical equations to characterize Spanish olive oils by varieties. *Grasas Aceites* **1988**, *39*, 219−228.
(10) Aparicio, R.; Ferreiro, L.; Cert, A.; Lanzon, A. Characterization of Andalusian Virgin olive oil. *Grasas Aceites* **1990**, *41*, 23−39.
(11) Aparicio, R.; Calvente, J. J.; Morales, M. T. Sensory authentication of European extra-virgin olive oil varieties by mathematical procedures. *J. Sci. Food Agric.* **1996**, *72*, 435−439.
(12) Tsimidou, M.; Macrae, R.; Wilson, I. Authentication of virgin olive oils using principal components analysis of triglyceride and fatty acid profiles: part 1. Classification of Greek olive oils. *Food Chem.* **1987**, *25*, 227−239.
(13) Defernez, M.; Kemsley, E. K.; Wilson, R. H. Use of infrared spectroscopy and chemiometrics for the authentication of fruit purees, *J. Agric. Food Chem.* **1996**, *44*, 175−180.
(14) Lai, Y. W.; Kemsley, E. K.; Wilson, R. H. Potential of Fourier transform infrared spectroscopy for the authentication of vegetable oils. *J. Agric. Food Chem.* **1994**, *42*, 1154−1159.
(15) Vogels, J. T. W. E.; Terwel, L.; Tas, A. C.; Van den Berg, F.; Dukel, F.; Van der Greef, G. Detection of adulteration in orange juice by a new screening method using proton NMR spectroscopy in combination with pattern recognition tecniques. *J. Agric. Food Chem.* **1996**, *44*, 175−180.
(16) Mannina, L.; Barone, P.; Patumi, M.; Fiordiponti, P.; Emanuele, M. C.; Segre, A. L. Cultivar and pedoclimatic effect in the discrimination of olive oils: A high-field NMR study. *Recent Res. Dev. Oil Chem.* **1999**, *3*, 85−92.
(17) Fauhl, C.; Reniero, F.; Guillou, C. ^{1}H NMR as a tool for the analysis of mixtures of virgin olive oils with oils of different botanical origin. *Magn. Reson. Chem.* **2000**, *38*, 436−443.
(18) Ng, S. Analysis of positional distribution of fatty acids in palm oil by ^{13}C NMR spectroscopy. *Lipids* **1985**, *20*, 778−782.
(19) Wollenberg, K. F. Quantitative high resolution ^{13}C nuclear magnetic resonance of the olefinic and carbonyl carbons of edible vegetable oils. *J. Am. Oil Chem. Soc.* **1990**, *67*, 487−494.
(20) Mannina, L.; Luchinat, C.; Emanuele, M. C.; Segre, A. L. Acyl positional distribution of glycerol triesters in vegetable oils: a ^{13}C NMR study. *Chem. Phys. Lipids* **1999**, *103*, 47−55.
(21) Mannina, L.; Luchinat, C.; Patumi, M.; Emanuele, M. C.; Rossi, E.; Segre, A. L. Concentration dependence of ^{13}C NMR spectra of triglycerides: implications for the NMR analysis of olive oils. *Magn. Reson. Chem.* **2000**, *38*, 886−890.
(22) Sacchi, R.; Patumi, M.; Fontanazza, G.; Barone, P.; Fiordiponti, P.; Mannina, L.; Rossi, E.; Segre, A. L. A high-field ^{1}H nuclear magnetic resonance study of the minor components in virgin olive oils. *J. Am. Oil Chem. Soc.* **1996**, *23*, 747−758.
(23) Sacchi, R.; Mannina, L.; Fiordiponti, P.; Barone, P.; Paolillo, L.; Patumi, M.; Segre A. L. Geographical classification of Italian extra-virgin olive oils by high field ^{1}H NMR spectroscopy. *J. Agric. Food Chem.* **1998**, *46*, 3947−3951.
(24) Segre, A. L.; Mannina, L. ^{1}H NMR study of edible oils. *Recent Res. Dev. Oil Chem.* **1997**, *1*, 297−308.
(25) Mannina, L.; Patumi, M.; Fiordiponti, P.; Emanuele, M. C.; Segre, A. L. Olive and hazelnut oils: a study by high-field ^{1}H NMR and gas chromatography. *Ital. J. Food Sci.* **1999**, *11*, 139−149.
(26) Mannina, L.; Patumi, M.; Proietti, N.; Segre, A. L. D.O.P. (Denomination of Protected Origin) geographical characterization of Tuscan extra virgin olive oils using high-field ^{1}H NMR spectroscopy. *Ital. J. Food Sci.* **2000**, in press.
(27) Morrison D. F. *Multivariate Statistical Methods*, 3rd ed.; McGraw-Hill: New York, 1990; p 269.
(28) Romesburg, H. C. *Cluster Analysis for Researchers*; Krieger Publishing: Malabar, FL, 1984.

Received for review November 22, 2000. Revised manuscript received February 27, 2001. Accepted March 5, 2001. This work has been partly financed by a contribution from the EEC in the Promotion of Oliviculture program Ricerca Sviluppo Tecnologico e Alta Formazione.

JF001408I