JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Geographical Characterization of Greek Virgin Olive Oils (Cv. Koroneiki) Using ^{1}H and ^{31}P NMR Fingerprinting with Canonical Discriminant Analysis and Classification Binary Trees

PANOS V. PETRAKIS,*,† ALEXIA AGIOMYRGIANAKI,§ STELLA CHRISTOPHORIDOU,§
APOSTOLOS SPYROS,§ AND PHOTIS DAIS§

Laboratory of Entomology, Institute of Mediterranean Forest Research, National Agricultural Research Foundation, Terma Alkmanos, Ilissia, 115 28 Athens, Greece, and NMR Laboratory, Department of Chemistry, University of Crete, P.O. Box 2208, Voutes campus, 710 03 Heraklion, Crete, Greece

This work deals with the prediction of the geographical origin of monovarietal virgin olive oil (cv. Koroneiki) samples from three regions of southern Greece, namely, Peloponnesus, Crete, and Zakynthos, and collected in five harvesting years (2001−2006). All samples were chemically analyzed by means of ^{1}H and ^{31}P NMR spectroscopy and characterized according to their content in fatty acids, phenolics, diacylglycerols, total free sterols, free acidity, and iodine number. Biostatistical analysis showed that the fruiting pattern of the olive tree complicates the geographical separation of oil samples and the selection of significant chemical compounds. In this way the inclusion of the harvesting year improved the classification of samples, but increased the dimensionality of the data. Discriminant analysis showed that the geographical prediction at the level of three regions is very high (87%) and becomes (74%) when we pass to the thinner level of six sites (Chania, Sitia, and Heraklion in Crete; Lakonia and Messinia in Peloponnesus; Zakynthos). The use of classification and binary trees made possible the construction of a geographical prediction algorithm for unknown samples in a self-improvement fashion, which can be readily extended to other varieties and areas.

KEYWORDS: Fatty acids; phenolics; NMR spectroscopy; virgin olive oil; discriminant analysis; classification binary trees; chemometrics; geographical prediction

INTRODUCTION

Virgin olive oil constitutes a valuable commodity of Greece, which is the third major producer in the Mediterranean basin at ∼400,000 tons per year. Crete, the largest island of Greece in the Mediterranean Sea, and the southern part of Peloponnesus, especially the seaside plains of Lakonia and Messinia, are characterized by a mild climate and fertile soil, creating thus ideal conditions for the cultivation of olive trees and the production of olive oil. There are about 10 prominent varieties for the production of olive oil: Koroneiki, the best known of all, is cultivated in several areas. It is the strongest of the Greek cultivars and the least demanding in terms of moisture, soil, and care. Koroneiki is a tree of medium vigor with a spreading habit and an open canopy (*1*). Its oil is considered to be the finest of all other varieties, endowed with exceptional aroma and taste. The fruit ripens early and has very high oil content. The produced olive oil is high in oleic acid (74−80%), total polar phenols (200−350 mg/kg), and α-tocopherol (13−75 mg/kg), which are responsible for its high stability against oxidation and its nutritional−therapeutic properties (*2*). These quality characteristics of olive oils extracted from cv. Koroneiki allow the labeling of olive oils as Protected Designation of Origin (PDO) and/or Protected Geographical Indication (PGI) products (provided that a well-defined registration process is completed on the basis of administrative type of data. This designation guarantees that the quality of the product is closely linked to its territorial origin.

In the past decade, there have been intensive studies for the determination of the geographical origin of olive oils using varied physical and/or chemical parameters in combination with chemometrics. There are two main directions in these studies: in the first one, multivariate statistical methods were applied to compositional parameters obtained by analytical methods, such as gas chromatography (GC), high-performance liquid chromatography (HPLC), and mass spectrometry coupled or not with GC and HPLC. Discrimination of olive oil samples from different geographical origins has been achieved on the basis of different olive oil constituents, including fatty acids and triacylglycerols (*3*−*6*), sterols (*7*), volatiles (*8*−*10*), and trace elements (*11*). The second direction involves spectroscopic

* Corresponding author (telephone +302107790865; fax +30210-7784602; e-mail pvpetrakis@fria.gr).
† National Agricultural Research Foundation.
§ University of Crete.

10.1021/jf072957s CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/19/2008

studies on samples without previous treatment either using the whole spectrum or specific peak intensities and/or chemical shifts. In this direction belong classification studies employing ^{1}H and ^{13}C NMR spectroscopy (*12−16*), near-infrared spectroscopy (NIR) (*17−19*), Fourier transformed infrared spectroscopy (FTIR) (*20*), Fourier transformed Raman spectroscopy (*21*), and fluorescence spectroscopy (*22*).

Among the various compositional parameters used for the discrimination of olive oils fatty acids appear to be extremely useful, because their composition and positional distribution in triacylglycerols are affected by a number of different factors, such as genotype, ripening stage, and pedoclimatic conditions (*3−6, 23*). Moreover, minor compounds, such as polyphenols and diacylglycerols, are very sensitive to olive fruit cultivar, maturity stage, geographical origin, and agronomic practices and therefore were used to discriminate olive oil cultivars (*24−26*). Their ability to predict the discrimination of olive oils' geographical location has not been investigated yet.

The present work aims at the classification of 131 EVOO samples produced from cv. Koroneiki by using solely NMR spectroscopy. The oil samples were collected from different areas of Crete and Peloponnesus and from the island of Zakynthos. The regions of Messinia and Lakonia were chosen from Peloponnesus, which is considered to be the native place of Koroneiki. Three sites from the island of Crete (Sitia, Chania, and Heraklion) were selected, because it has been an important olive oil producing area since Minoan times. In these two regions, there are 13 PDO and 2 PGI olive oils. The olive oils produced outside these regions, including Zakynthos, are PGI oils (*27*). Classification will be performed on different compositional parameters (variables) determined by ^{1}H NMR (fatty acids, iodine number) (*28*) and ^{31}P NMR (phenolic compounds, diacylglycerols, total free sterols, and free acidity) (*29*). Instead of peak intensities, we preferred a targeted analysis and use chemical components defined a priori, because their identity allows interpretation of the results with biological and genetic terms, when possible. At first, the olive oils were classified according to the three major geographic units (Crete, Peloponnesus, and Zakynthos), then within the geographical units, and finally the classification criterion is augmented by the harvesting year.

MATERIALS AND METHODS

EVOO Samples. A total of 131 EVOO samples were obtained from three regions of Crete (15 samples from Sitia, 34 from Heraklion, and 25 from Chania), two regions of Peloponnesus (29 samples from Messinia and 9 from Lakonia), and 19 samples from the island of Zakynthos. The virgin olive oil samples were provided by the local cooperatives from small groups of neighboring olive trees within the same tenth hectare field area and produced by the same method of extraction (centrifugation). Virgin olive oils were extracted within 48 h after harvesting, stored immediately at −20 °C, and analyzed twice within 2 months from the production date. The virgin olive samples considered in this study came from small groups of neighboring olive trees of cv. Koroneiki within the same tenth hectare field area. By using only one variety we expect to reduce the variability imposed if many varieties were selected. Cv. Koroneiki was chosen because it is a very old variety, and it is expected that time was sufficient for the action of several biological processes, such as the achievement of genetic integrity by reducing gene flow (*1*). The oil samples were collected during five consecutive harvesting periods 2001−2006. The date of extraction of the olive oils was different in the various locations ranging from November to January. All of the Greek olive oils were virgin according to certain official analytical methods and limits (free acidity $\leq$ 0.8% in oleic acid, $K_{232} \leq 2.50$, $K_{270} \leq 0.20$, $\Delta K \leq 0.01$, total sterols $\geq$ 1000 mg/kg) (*30*). There is no doubt that the greater the number of samples, the more secure the analysis in the sense that the entire variability of the subject population is included in the sample. Admittedly, the representation of harvesting years is not homogeneous. However, this does not confer any mathematical problem to the employed mathematical techniques (see below). There is no rule saying that some samples may be repeated, and in this sense there is no bias in these techniques.

Pinacol, triethylamine, phosphorus trichloride, protonated solvents (reagent or analytical grade), and deuterated solvents used in the present study were purchased from Sigma-Aldrich (Athens, Greece). The derivatizing phosphorus reagent [2-chloro-4,4,5,5-tetramethyldioxaphospholane (**1**)] was synthesized from pinacol and phosphorus trichloride following the method described in the literature (*31*). However, to increase the yield of the reaction, we utilized hexane solvent instead of benzene and pyridine instead of triethylamine as suggested in the original method. This modification resulted in ~ 45% yield of the product against 19% obtained with the original method.

Extraction of Minor Polar Components from Olive Oil. Phenolic compounds were extracted following the method developed by Montendoro et al. (*32*) using 35 g of olive oil and a mixture of ethanol/water (80:20, v/v). The polyphenol extracts were used immediately for sample preparation and ^{31}P NMR measurements.

Sample Preparation for ^{31}P NMR Spectral Analysis. A stock solution was prepared by dissolving 0.6 mg of chromium acetylacetonate [$Cr(acac)_3$] (0.165 μM) and 13.5 mg of cyclohexanol (13.47 mM) in 10 mL of a mixture of pyridine and $CDCl_3$ solvents (1.6:1.0 volume ratio) and protected from moisture with 5A molecular sieves. Cyclohexanol was used as an internal standard for quantification purposes. A small quantity of olive oil (100−150 mg) was placed in a 5 mm NMR tube. The required volumes of the stock solution (0.4 mL) and reagent **1** (15 μL) were added. The reaction mixture was left to react for about 15 min at room temperature. Upon completion of the reaction, the solution was used to obtain the ^{31}P NMR spectra.

NMR Experiments. All NMR experiments were conducted on a Bruker AMX500 spectrometer operating at 500.1 and 202.2 MHz for proton and phosphorus-31 nuclei, respectively, at 26 ± 1 °C.

One-dimensional ^{13}P NMR spectra were recorded by employing the inverse gated decoupling technique to suppress NOE effects. Typical spectral parameters for quantitative studies using reagent **1** were as follows: 90° pulse width = 12.5 μs; sweep width = 55 kHz; relaxation delay = 25 s; memory size = 32 K. To ensure quantitative spectra, the magnitude of the relaxation delay adopted was >5 times the relaxation time (T_1 = 4.6 s) of the phosphitylated cyclohexanol; 32 transients were accumulated for each spectrum. For all FIDs, line broadening of 1 Hz was applied and drift correction was performed prior to Fourier transform. Polynomial fourth-order baseline correction was performed before integration.

One-dimensional high-resolution ^{1}H NMR spectra were acquired with the following acquisition parameters: time domain = 32 K; 90° pulse width = 9.3 μs; spectral width = 12 ppm; relaxation delay = 2 s; 32 scans and 8 dummy scans were accumulated. Base-line correction was performed carefully by applying a polynomial fourth-order function to achieve a quantitative evaluation of all signals of interest. The spectra were acquired without spinning the NMR tube in order to avoid artificial signals, such as spinning side bands of the first or higher order.

Biostatistical Methods. Canonical discriminant analysis (CDA) is used to achieve the most discriminative variables for the arrangement of samples in a space of reduced dimensionality in a way that maximizes the distances between the a priori formed groups and the independence of the axes of the configuration (*33, 34*). The EVOO chemical compounds contributing most to the discrimination of groups are shown by means of the *F* value used as a criterion for inclusion or removal of the compound in a forward stepwise CDA mode. Wilks' λ and the associated *F* approximation were used to check the significance and estimate the importance of each compound in CDA. The classification tables of CDA are produced to present the classificatory efficiency of the method rather than illustrating the group relations of the samples. On the basis of the resulting chemical profile the Mahalanobis distances were calculated in estimating the misclassification probabilities of EVOO samples in the predefined groups. Two CDAs are used for the

Table 1. Compositional Parameters of EVOO Samples from Six Sites in Greece: Means, Standard Deviation, Determined by NMR

	Chania, $N^a = 25$		Heraklion, $N = 34$		Lakonia, $N = 9$		Messinia, $N = 29$		Sitia, $N = 15$		Zakynthos, $N = 19$	
	mean	±sd	mean	±sd	mean	±sd	mean	±sd	mean	±sd	mean	±sd
syringaresinol	1.31	0.55	0.98	0.39	0.81	0.31	0.86	0.30	1.14	0.25	0.69	0.18
1-acetoxypinoresinol	3.82	1.33	3.04	1.32	2.93	0.83	3.13	0.55	2.88	0.86	3.16	0.64
t-hydroxytyrosol[b]	51.23	37.58	80.82	70.01	52.24	40.79	42.09	21.98	134.22	104.80	43.04	16.71
t-tyrosol	42.18	19.66	59.58	33.18	42.59	20.27	43.56	14.53	80.71	54.83	55.54	13.40
p-coumaric acid	0.16	0.17	0.15	0.11	0.20	0.14	0.15	0.14	0.21	0.16	0.13	0.16
f-tyrosol	4.50	2.12	6.48	8.07	2.97	1.40	3.52	2.28	5.31	2.62	3.21	3.61
f-hydroxytyrosol	3.28	3.15	4.16	3.38	3.68	1.45	2.06	1.66	6.24	3.89	0.92	0.72
homovanillyl alcohol	3.02	1.81	3.39	1.31	1.94	0.96	2.58	1.59	3.93	4.06	2.25	1.86
luteolin	0.57	0.32	0.63	0.48	0.41	0.26	0.39	0.30	0.71	0.29	0.12	0.06
apigenin	0.26	0.17	0.45	0.43	0.18	0.14	0.22	0.16	0.23	0.08	0.11	0.06
pinoresinol	1.76	1.44	1.27	1.98	1.55	0.75	1.88	1.70	1.60	1.78	2.42	1.60
linolenic acid	0.46	0.11	0.47	0.13	0.42	0.14	0.38	0.15	0.47	0.12	0.39	0.16
linoleic acid	6.44	1.49	7.32	1.75	5.87	1.13	5.85	1.23	8.42	1.66	4.42	1.08
oleic acid	77.63	2.69	77.18	3.12	76.95	3.05	80.27	3.45	76.03	3.03	77.91	3.16
saturated fatty acids	15.47	1.90	15.03	1.68	16.76	3.29	13.50	2.97	15.08	1.69	17.29	3.42
iodine value	78.38	1.60	79.74	1.83	77.44	3.44	80.13	2.68	80.53	1.36	75.66	3.50
1,2-diacylglycerols	1.69	0.39	1.74	0.33	1.85	0.26	1.75	0.30	1.78	0.21	1.72	0.33
1,3-diacylglycerols	0.28	0.15	0.28	0.16	0.16	0.15	0.24	0.17	0.27	0.24	0.17	0.11
total diacylglycerols	1.98	0.48	2.02	0.35	2.01	0.26	1.99	0.39	2.04	0.20	1.89	0.40
D ratio	0.86	0.06	0.86	0.07	0.92	0.07	0.89	0.06	0.87	0.10	0.92	0.04
free acidity	0.35	0.15	0.33	0.15	0.28	0.13	0.21	0.16	0.29	0.09	0.21	0.12
sterols	0.10	0.02	0.11	0.03	0.09	0.03	0.09	0.02	0.11	0.02	0.08	0.02

[a] Number of samples. [b] "t-" stands for total and "f-" for free.

Table 2. Summary of the Steps for the Interactive Forward Mode of CDA[a]

	CDA for three geographical divisions				CDA for six geographical sites			
compound	Wilks' λ[b]	approx *F* value; df_1, df_2	*P*[c]	rank in harvest year[d]	Wilks' λ	approx *F* value; df_1, df_2	*P*	rank in harvest year
linoleic acid	0.686	29.27; 2, 128	*	1	0.602	16.52; 5, 125	*	1
oleic acid	0.584	19.57; 4, 254	*	5	0.457	11.88; 10, 248	*	7
pinoresinol	0.504	17.17; 6, 252	*	19	0.144	7.23; 40, 517	*	
1,2-diacylglycerols	0.448	15.42; 8, 250	*		0.193	8.16; 30, 482	*	15
acidity	0.398	14.49; 10, 248	*	2				
f-hydroxytyrosol[e]	0.359	13.72; 12, 246	*	20	0.111	6.60; 50, 532	*	20
t-tyrosol[e]	0.321	13.35; 14, 244	*	11	0.164	7.72; 35,503	*	11
p-coumaric acid	0.289	12.99; 16, 242	*	14				
t-hydroxytyrosol	0.275	12.07; 18, 240	*	15	0.128	6.81; 45, 526	*	13
apigenin	0.240	12.39; 20, 238	*	12	0.356	10.26; 15, 339	*	5
iodine value	0.231	11.58; 22, 236	*	6				
luteolin	0.222	10.94; 24, 234	*	17				
syringaresinol					0.285	9.31; 20, 405	*	12
acetoxypinoresinol					0.234	8.61; 25, 450	*	9

[a] The most important variables seen in the two CDAs for three geographical divisions and six sites of origin. [b] The compounds are sorted according to their Wilks λ significance. [c] Significance at a *P* value of $<10^{-4}$. [d] Rank of the compounds in the CDAs when the harvesting year is incorporated in the definition of groups. [e] "t-" stands for total and "f-" for free.

groupings according to two hierarchical levels, namely, geographical units and individual sites within each unit area. In both levels the performance of CDA was measured by the percentage of total variance explained by all significant discriminant axes.

Classification binary trees (CBTs) are used to produce a set of simple rules, which will identify the origin of any new sample (*35, 36*). The construction of a CBT comprises the split of the original set of samples into two parts on the basis of a criterion involving a few variables, usually an algebraic expression of one or two. All variables involved in the construction of a CBT comprise the best diagnostic variable set that can predict the group affiliation of samples. The data are classified here into the "leaves" of the tree; in an ideal situation six site groups are produced. This noise causes the reduction of significant variables in the finer geographical scale (**Table 2**) because fewer compounds conform and contribute to the existing separation of sites. The rationale of constructing a framework for the group affiliation of existing EVOO samples is that a set of simple rules predicts the origin of the already existing and future EVOO samples of unknown origin. The group identifier was used as a dependent variable, whereas independent variables are derived by NMR (**Table 1**). The reduction of error in the classification is monitored by means of a loss function. Several loss functions have been proposed (*36*), and in this study the fitting method was the *Gini* index on the basis of the better reduction in error it achieved.

The methods CDA and CBT employed in the analysis of the EVOO samples are preferred over other methods pertaining to the matter of geographical prediction of origin because they possess several features: (a) CDA provides a statistical test for the intrinsic dimensionality in the data; (b) both CDA and CBT estimate the importance of the variables, that is, chemical compounds, whereas CDA provides additional statistical tests of their significance; (c) unlike other ordination methods, they start from known cluster affiliation and test their integrity according to the describing compounds; (d) the outcome, especially of CBT analysis, can be easily implemented in procedural and easily understandable computer algorithms; (e) they can be cross-checked and handle many compounds and samples (*34*), and for this they can be used in the construction of artificial intelligence systems.

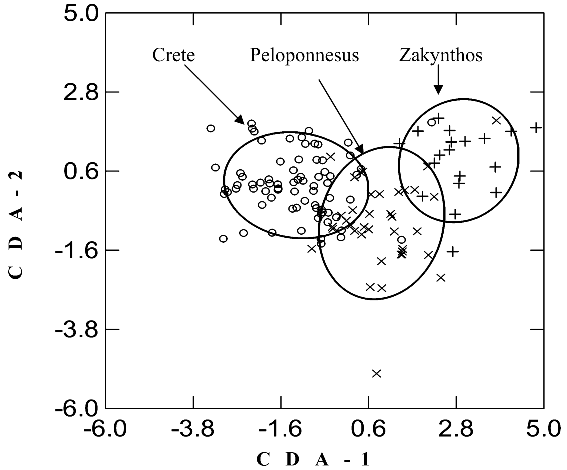

Figure 1. CDA scattergram of the three geographical divisions of Greece. The axes (CDA-1 = 87.8% and CDA-2 = 12.2%) account for the total variability of the measured variables. Ninety-five percent ellipses are drawn around each centroid of groupings in such a way that leaves outside the misclassified olive oil samples. The positions of the groups are perfectly correlated with their geographical arrangement.

RESULTS AND DISCUSSION

The mean values and standard deviations of the compositional parameters used in this study for the 131 EVOO samples as a whole are summarized in **Table 1**. The results for each individual sample are available as Supporting Information. **Table 1** contains data of fatty acids and iodine number obtained by ^{1}H NMR spectroscopy, whereas **Table 2** depicts the data of the two diacylglycerol isomers (1,2-DGs and 1,3-DGs), the ratio D (1,2-DGs/total DGs), total free sterols, free acidity, and phenolic compounds obtained by employing ^{31}P NMR spectroscopy. Spectral assignments and methods of quantification of the NMR data have been reported in detail in previous publications (*28, 29*).

The CDA of geographical divisions, that is, Peloponnesus, Crete, and Zakynthos, and the six sites of origin are depicted in **Figure 2**. The relative positions of groups reflect their geographical locations. In **Figure 1** the first axis (CDA-1) bears the bulk of the variation in the data (87.8%), whereas the second axis (CDA-2) bears 12.2% variation if the original data do not contribute significantly to the discrimination of divisions. Both axes are drawn to show the entire variation (100%) in a highly significant analysis (Wilks' $\lambda = 0.24$, $F_{\text{approx}} = 8.68$, $\text{df}_1 = 28$, $\text{df}_2 = 230$, $P < 10^{-4}$). The compounds that are statistically responsible are shown in **Table 2**. For the initial variables set of 37 compounds, 12 are found to be statistically highly significant for this particular CDA. Eight of them can be also found in the section **B** of the same table, which depicts the variables set for the six sites of origin. Among variables there is a strong tendency to co-occur in organisms because they reflect the presence of a certain genetic background, which is responsible for the phenolic profile of the samples given their common extraction procedure within 2 months from harvest. This implies the presence of the same or similar biogenetic enzyme(s) and the absence or low activity of the enzyme(s) that guide the catabolism of these compounds (*45*). If these enzymes are present, it is supposed that the particular abiotic conditions of an area have a specific environmental effect on the activities of these enzymes and the substrate on which they are acting, whereas the plant genetic substrate is the same because of the same olive tree group.

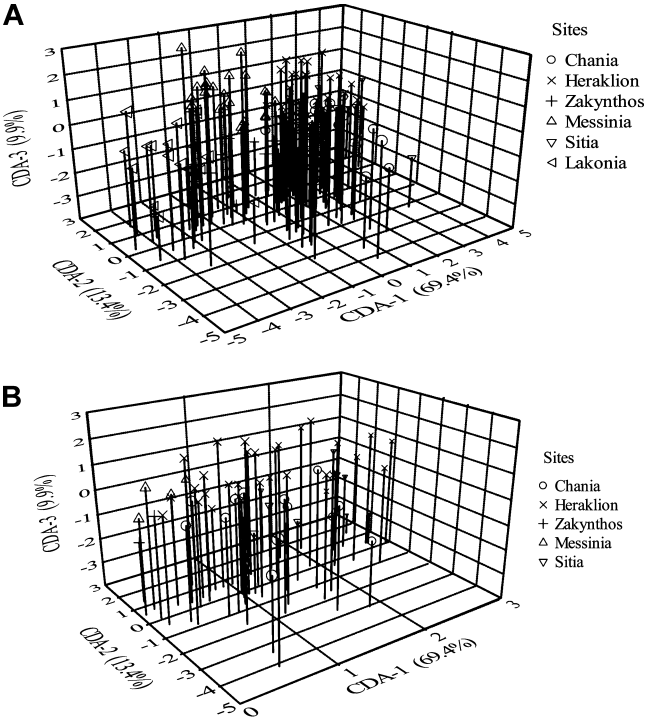

Figure 2. Relative positions in a CDA three-dimensional space of the EVOO samples from the six sites in Greece. The axes (CDA-1 = 69.4%, CDA-2 = 13.4%, and CDA-3 = 9.9%) account for 92.6% of the variability in the measured variable set. (**A**) shows the entire set of data, and (**B**) illustrates the central cluster of samples having CDA-1 scores between 0 and 3.

When CDA is applied to the grouping of the six sites, the analysis is strongly significant (Wilks' $\lambda = 0.065$, $F_{\text{approx}} = 5.94$, $\text{df}_1 = 70$, $\text{df}_2 = 537$, $P < 10^{-4}$) and the significant CDA axes are three, accounting for 69.4, 13.4, and 9.9%, amounting to a total 92.6% of the existing variation. The most significant compounds for this CDA are 10, 2 fewer than the previous CDA. This observation is strange at first glance because the six sites are at a finer level of geographical partition, and it is expected to be differentiated at a finer scale involving more compounds. This discrepancy shows that there must be considerable noise in the data set introduced by the harvesting year, because the olive tree in Greece exhibits an every-other-year fruiting pattern. In this respect, the classification efficiency of the aforementioned CDAs summarized in **Tables 3A**,**B** is the result of both different geographical levels and noisy data.

The existence of similar compounds in the set of significant compounds is another salient feature of **Table 2**. In this table it is shown that 80% of the compounds are common between the two CDAs. This means that the separation of sites within larger areas is based on quantitative differences and alteration of very similar compounds. An example is the pair of recently discovered lignans, pinoresinol (CDA of three divisions) and 1-acetoxypinoresinol (CDA of six sites), which in the olive oil are presumably synthesized via the same biogenetic path (*46*).

Figure 2 shows the discriminant arrangement of the geographical origins of the EVOO samples from the six sites in Greece. Although the correlation with geographical positions is evident, zero lines of the axes separate individual samples. A representative example is the site Chania, which is separated by the zero line on the first discriminant axis (CDA-1), whereas Zakynthos is separated by the same lines on CDA-2 and CDA-3. Sites on Crete are separated by all axes, yet they retain their geographical positions. This indicates that the relationship

Table 3. Classification Table of the CDA of EVOO Samples Grouped According to (**A**) Three Divisions, (**B**) Six Sites of Origin, and (**C**) Six Sites of Origin Taking into Account the Harvesting Year[a]

(**A**) Three Divisions

	Crete	Peloponnesus	Zakynthos
Crete	*64*	9	1
Peloponnesus	2	*33*	3
Zakynthos	0	2	*17*
total	**66**	**44**	**21**

(**B**) Six Sites of Origin

	Chania	Heraklion	Lakonia	Messinia	Sitia	Zakynthos	% correct classification
Chania	20	2	1	1	0	1	80 (56)
Heraklion	3	23	1	2	5	0	68 (56)
Lakonia	0	1	6	1	0	1	67 (33)
Messinia	2	4	2	19	0	2	66 (52)
Sitia	1	3	0	0	11	0	73 (60)
Zakynthos	0	0	0	1	0	18	95 (89)
total	**26**	**33**	**10**	**24**	**16**	**22**	**74 (59)**

(**C**) Six Sites of Origin, Taking into Account Harvest Year

harvest year	site	12Chania	12Herakli	12Messini	12Sitia	23Chania	23Herakli	23Lakonia	23Sitia	34Chania	34Lakonia
2001–2002	Chania	9	1	0	0	0	0	0	0	0	0
2001–2002	Heraklion	1	13	0	1	0	0	0	0	0	0
2001–2002	Messinia	0	0	4	0	0	0	0	0	0	0
2001–2002	Sitia	0	3	0	6	0	0	0	0	0	0
2002–2003	Chania	0	0	0	0	5	0	0	0	0	0
2002–2003	Heraklion	0	0	0	0	0	12	0	2	0	0
2002–2003	Lakonia	0	0	0	0	0	0	1	0	0	0
2002–2003	Sitia	0	0	0	0	0	0	0	6	0	0
2003–2004	Chania	0	0	0	0	0	0	0	0	4	0
2003–2004	Lakonia	0	0	0	0	0	0	0	0	0	1
2003–2004	Messinia	0	0	0	0	0	0	0	0	0	0
2004–2005	Heraklion	0	0	0	0	0	0	0	0	0	0
2004–2005	Lakonia	0	0	0	0	0	0	0	0	0	0
2004–2005	Zakynthos	0	0	0	0	0	0	0	0	0	0
2005–2006	Chania	0	0	0	0	0	0	0	0	0	0
2005–2006	Heraklion	0	0	0	0	0	0	0	0	0	0
2005–2006	Lakonia	0	0	0	0	0	0	0	0	0	0
2005–2006	Messinia	0	0	0	0	0	0	0	0	0	0
2005–2006	Zakynthos	0	0	0	0	0	0	0	0	0	0
total		10	17	4	7	5	12	1	8	4	1

harvest year	site	34Messini	45Herakli	45Lakonia	45Zakynth	56Chania	56Herakli	56Lakonia	56Messini	56Zakynth	% correct
2001–2002	Chania	0	0	0	0	0	0	0	0	0	90 (80)
2001–2002	Heraklion	0	0	0	0	0	0	0	0	0	87 (33)
2001–2002	Messinia	0	0	0	0	0	0	0	0	0	100 (0)
2001–2002	Sitia	0	0	0	0	0	0	0	0	0	67 (44)
2002–2003	Chania	0	0	0	0	0	0	0	0	0	100 (40)
2002–2003	Heraklion	0	0	0	0	0	0	0	0	0	86 (71)
2002–2003	Lakonia	0	0	0	0	0	0	0	0	0	100 (0)
2002–2003	Sitia	0	0	0	0	0	0	0	0	0	100 (100)
2003–2004	Chania	0	0	0	0	0	0	0	0	0	100 (100)
2003–2004	Lakonia	1	0	0	0	0	0	0	0	0	50 (0)
2003–2004	Messinia	11	0	0	0	0	0	0	0	0	100 (82)
2004–2005	Heraklion	0	2	0	0	0	0	0	0	0	100 (100)
2004–2005	Lakonia	0	0	5	0	0	0	0	0	0	100 (40)
2004–2005	Zakynthos	0	0	0	6	0	0	0	0	0	100 (83)
2005–2006	Chania	0	0	0	0	5	0	0	0	1	83 (50)
2005–2006	Heraklion	0	0	0	0	0	3	0	0	0	100 (33)
2005–2006	Lakonia	0	0	0	0	0	0	1	0	0	100 (0)
2005–2006	Messinia	0	0	0	0	0	0	0	14	0	100 (71)
2005–2006	Zakynthos	0	0	0	0	0	0	0	0	13	100 (85)
total		12	2	5	6	5	3	1	14	44	92 (63)

[a] The first two digits in column labels indicate the harvesting year, e.g., 45Lakonia corresponds to samples originating from Laconia, harvested in 2004–2005). The first column includes the areas, which are predicted to belong to the areas of the first row. The concentration of values on the main diagonal denotes the appropriateness of the measured variables. The values in parentheses in the last column are from jack-knifing (leave-one-out) of CDA, which is the ability of the other regions to predict the correct category to which the sites belong.The designation "*xy*Chania" means the sample originated at Chania in the harvesting period 200*x*–200*y*.

between compounds and sites may be complex, but the outcome is an impressive coincidence between geographical and discriminant positions. The separation of Cretan sites from the other three is mainly caused by total (t-) and free (f-) hydroxytyrosol and t-tyrosol contents, which are represented with higher concentrations in Cretan samples than the rest (**Table 1**) and

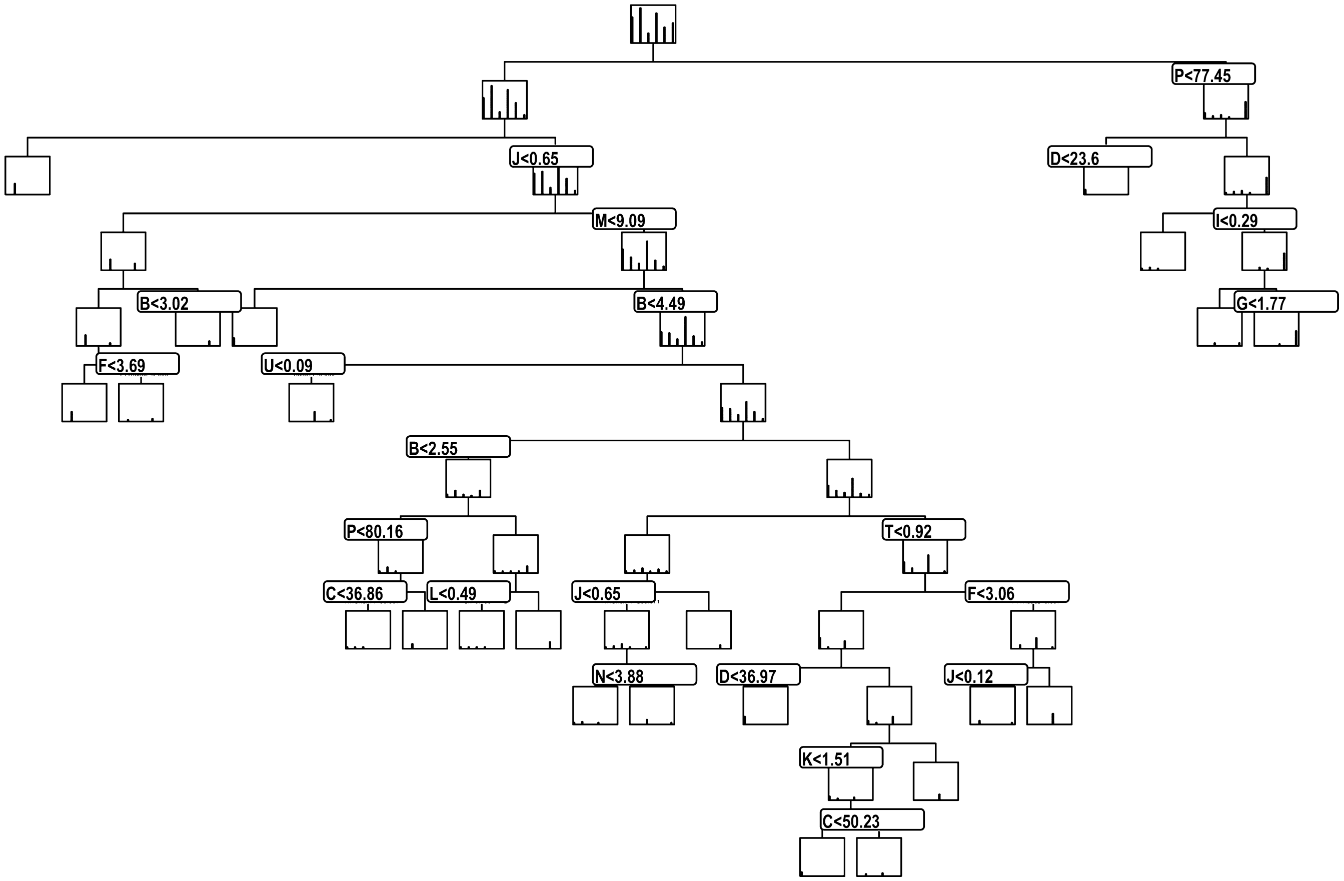

Figure 3. CBT of the EVOO samples from six geographical sites. The sites at the root mobile refer to Chania, Heraklion, Lakonia, Messinia, Sitia, and Zakynthos. In this order they are shown in the entire CBT tree. The original groups are indicated by dividing the base of the square in six segments, and the samples are small circles stacked in proportion to the size of each group. At each node the cluster is characterized by the simple inequality written involving one compound in the rounded square. **Table 4** symbolizes compound names with one capital letter.

simultaneously are highly loaded by the standardized discriminant functions on the CDA-3. However, this axis explains only a small percentage of the total variation (9.9%).

In both CDAs the achieved classification is very good (**Table 3A**,**B**). Due to the aforementioned complexity the correct classifications amount to 87 and 74% of the EVOO samples in the two CDAs, respectively. The same table shows that the ability of the sites to be predicted by the position of the others is much reduced in section **B**, especially at Chania (in relation to the first classification, 80%) and Lakonia, which are dominated by isolated areas and extreme relief with many gorges and cliffs. The biotic result of this condition is the high levels of endemism of these sites (*38*). This supports the idea that phenylalanine ammonia-lyase enzyme (PAL), which catalyzes the biogenesis of phenolics, is subjected to important alterations due to abiotic and biotic factors (*39*); this seems to be the case at the sites at Chania and Lakonia.

When the harvesting year is taken into account in the formation of groups, then CDA is complex as expected. Therefore, some authors remove the factor "harvesting year" prior to statistical treatment. The significant eigenvalues are 43.25, 24.34, 8.54, and 6.37, accounting for 82.50% of the variation in the data; the set of 22 significant compounds was judged to be significant. This decrease is compensated by an increase in the classification efficiency, which now is 92% (**Table 3C**). Together with the fact that the ability of samples to be identified by the rest is reduced, this CDA implies that the fruiting pattern of the olive tree complicates the ability of CDA. There is no clear trend in EVOO samples implying that the prediction of the site of origin must be deduced from harvesting in the same year. This complication is added to the intrinsic high polymorphism within *Olea europaea*. Indeed, independent sites using differential staining of chromatin (*40*) showed that *O. europaea* cultivars impose a differentiation which is reflected to the chromosome polymorphism of olive trees and is imprinted in observable marks on chromosomes I, V, and VII of $2n = 46$. The observed polymorphism could not be compensated by gene flow among neighboring olive groves. The EVOO samples used here are extracted from the very old variety cv. Koroneiki. The extensive gene flow is expected to affect this cultivar at a higher degree than newer cultivars. However, our data show that this did not happen in the case of the three divisions or six sites. This contradicts the results of other workers who found, on the basis of morphological data, that the produced geographic pattern in the East Mediterranean is totally meaningless (*44*). The same conclusion was reached by Loukas and Krimbas (*1*), who examined the origins of 22 Greek varieties on the basis of allozyme markers.

Without regard to the classification ability of CDA, there is a need for a scheme to identify samples in a predefined algorithm. This is achieved by the straightforward classification of CBTs presented as a *mobile* (*35*) in **Figure 3**. The compounds at the nodes of the mobile as seen in **Table 4** do not coincide in terms of meaning and importance to those of **Table 2**. They are simply the criteria for the samples of the respective branch. The overall performance of the mobile structure to describe the site grouping is strong because the proportional reduction in the error is $r = 0.80$. Despite the very high reduction in error, due to the complex role of the compounds, the correspondence of the predefined groups to the leaves of the mobile is far from

Table 4. Compounds Participating in the Construction of CBT (Overall Reduction Error is 0.80)

symbol	important compound at node[a]	proportional reduction of error	% improvement[b]
P	iodine value	0.091	0.091
D	t-tyrosol	0.137	0.046
I	luteolin	0.17	0.033
F	f-hydroxytyrosol	0.183	0.013
J	apigenin	0.247	0.064
M	linoleic acid	0.308	0.062
B	1-acetoxypinoresinol	0.348	0.039
F	f-tyrosol	0.367	0.019
B	1-acetoxypinoresinol	0.419	0.053
U	acidity	0.457	0.038
B	1-acetoxypinoresinol	0.489	0.031
P	iodine value	0.518	0.029
L	linolenic acid	0.546	0.028
C	t-hydroxytyrosol	0.557	0.011
T	D ratio	0.593	0.036
C	t-hydroxytyrosol	0.62	0.027
N	homovanillyl alcohol	0.642	0.022
F	f-tyrosol	0.675	0.033
J	apigenin	0.719	0.044
D	t-tyrosol	0.757	0.038
K	pinoresinol	0.776	0.019
C	t-hydroxytyrosol	0.798	0.022

[a] A compound may appear in more than one node. [b] Describes the percentage reduction in error when the respective compound enters in the analysis.

perfect. In the mobile in **Figure 3** it is evident that there is no simple way to classify the samples from a single cultivar of *O. europaea*. Nevertheless, an unknown sample can be always classified in a leaf of the mobile. In conjunction with the classification produced by CDA, it can much reduce the misclassification error. In other studies (*41*) the CBT was used as the only means to classify unknown EVOO samples.

The inclusion of both linoleic and oleic acids in the significant variable set for the discrimination of EVOO geographic groups coincides with the finding that the polyunsaturated fatty acids in plants are not only important components of oil biochemistry but also good markers of geographical origin. The conversion of oleic to linoleic acid is governed by the two recently discovered genes *OeFDA2* and *OeFDA6* (*42*) that encode the key enzymes for this pathway. These enzymes seem to play a key role in many plants such as *Arabidopsis thaliana* (*43*), soybean, parsley, rape, peanut, sesame, cotton, sunflower, and spinach (*42*). In *O. europaea* these specific genes seem to have survived the extensive gene flow. This flow has been documented among *O. europaea* species and subspecies and among wild trees and cultivars. Especially in Mediterranean olive trees it is proposed that the gene flow obscures the phylogenetic signal but not the geographical one (*45*). In this study we found that the other significant variables in **Table 2** complement the signal carried by linoleic and oleic acids and produce a meaningful arrangement of EVOO origin sites.

In summary, this study has demonstrated that the compositional parameters obtained by NMR spectroscopy and analyzed statistically using CDA and CBT were successful in classifying EVOO samples from three different divisions and six different sites of Greece.

Supporting Information Available: Concentrations of phenolic compounds and fatty acids of extra virgin olive oils. This material is available free of charge via the Internet at http://pubs.acs.org.

LITERATURE CITED

(1) Loukas, M.; Krimbas, C. B. History of olive cultivars based on their genetic distances. *J. Hortic. Sci.* **1983**, *58*, 121–127.

(2) Psomiadou, E.; Karakostas, K. X.; Blecas, G.; Tsimidou, M. Z.; Boskou, D. Proposed parameters for monitoring quality of virgin olive oil (Koroneiki cv). *Eur. J. Lipid Sci. Technol.* **2003**, *105*, 403–408.

(3) Tsimidou, M.; Karakostas, K. X. Geographical classification of Greek virgin olive oil by non-parametric multivariate evaluation of fatty acid composition. *J. Sci. Food Agric.* **1993**, *62*, 253–257.

(4) 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.

(5) Olivier, D.; Artaud, J.; Pinatel, C.; Durbec, J. P.; Guerere, M. Differentiation of French virgin olive oils RDOs by sensory characteristics, fatty acid and triacylglycerol compositions and chemometrics. *Food Chem.* **2006**, *97*, 362–393.

(6) Di Bella, G.; Maisano, R.; La Reps, L.; Lo Turco, V.; Salvo, F.; Dugo, G. Statistical characterization of Sicilian olive oils from the Peloritana and Maghrebian zones according to the fatty acid profile. *J. Agric. Food Chem.* **2007**, *55*, 6568–6574.

(7) Leardi, R.; Paganuzzi, V. Characterization of the origin of extra virgin olive oils by chemometric methods applied to the sterols fraction. *Riv. Ital. Sostanze Grasse* **1987**, *64*, 131–136.

(8) Vichi, S.; Pizzale, L.; Conte, L. S.; Buxaderas, S.; Lopez-Tamames, E. Solid-phase microextraction in the analysis of virgin olive oil volatile fraction: Characterization of virgin olive oils from two distinct geographical areas of northern Italy. *J. Agric. Food Chem.* **2003**, *51*, 6572–6577.

(9) Zunin, P.; Boggia, R.; Salvadeo, P.; Evangelisti, F. Geographical traceability of West Liguria extra virgin olive oils by the analysis of volatile terpenoid hydrocarbons. *J. Chromatogr., A* **2005**, *1089*, 243–249.

(10) Cavaliere, B.; De Nino, A.; Hayet, F.; Lazez, A.; Macchione, B.; Moncef, C.; Perri, E.; Sindona, G.; Tagarelli, A. A metabolomic approach to the evaluation of the origin of extra virgin olive oil: a convenient statistical treatment of mass spectrometric analytical data. *J. Agric. Food Chem.* **2007**, *55*, 1454–1462.

(11) Benincasa, C.; Lewis, J.; Perri, E.; Sindona, G.; Tagarelli, A. Determination of trace elements in Italian virgin olive oils and their characterization according to geographical origin by statistical analysis. *Anal. Chim. Acta* **2007**, *585*, 366–370.

(12) 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.

(13) Sacchi, R.; Mannina, L.; Fiordiponti, P.; Barone, P.; Paolillo, L.; Patumi, M.; Segre, A. Characterization of Italian virgin olive oils using ^{1}H-NMR spectroscopy. *J. Agric. Food Chem.* **1998**, *46*, 3947–3951.

(14) Mannina, L.; Patumi, M.; Proietti, N.; Bassi, D.; Segre, A. Geographical characterization of Italian extra virgin olive oils using high-field ^{1}H-NMR spectroscopy. *J. Agric. Food Chem.* **2001**, *49*, 2688–2696.

(15) Vlahov, G.; Del Re, P.; Simone, N. Determination of geographical origin of olive oils using ^{13}C nuclear magnetic resonance spectroscopy. I. Classification of olive oils of the Puglia region with denomination of protected origin. *J. Agric. Food Chem.* **2003**, *51*, 5612–5615.

(16) Rezzi, S.; Axelson, D. E.; Héberger, K.; Reniero, F.; Mariani, C.; Guillou, C. Classification of olive oils using high throughput flow ^{1}H NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks. *Anal. Chim. Acta* **2005**, *552*, 13–24.

(17) Bertran, E.; Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Montolin, I. Near infrared spectrometry and pattern recognition as screening methods for the authentication of virgin olive oils of very close geographical origin. *J. Near Infrared Spectrosc.* **2000**, *8*, 45–52.

(18) Downey, G.; McIntyre, P.; Davies, A. N. Geographic classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near-infrared spectroscopic data. *Appl. Spectrosc.* **2003**, *57*, 158–163.

(19) Galtier, O.; Dupuy, N.; Le Dréau, Y.; Ollivier, D.; Pinatel, C.; Kister, J.; Artaud, J. Geographical origins and compositions of virgin olive oils determined by chemometric analysis of NIR spectra. *Anal. Chim. Acta* **2007**, *595*, 136–144.

(20) Tapp, H. S.; Defernez, M; Kemsley, E. K. FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *J. Agric. Food Chem.* **2003**, *51*, 6110–6115.

(21) Muik, B.; Lendl, B.; Molina-Diaz, A.; Ayora-Canada, M. J. Direct, reagent-free determination of free fatty acid content in olive oil and olives by Fourier transform Raman spectrometry. *Anal. Chim. Acta* **2003**, *487*, 211–226.

(22) Dupuy, N.; Le Dréau, Y.; Ollivier, D.; Artaud, J.; Pinatel, C.; Kister, J. Origin of French virgin olive oil registered designation of origins predicted by chemometric analysis of synchronous excitation-emission fluorescence spectra. *J. Agric. Food Chem.* **2005**, *53*, 9361–9368.

(23) Lanteri, S.; Armanino, C.; Perri, E.; Palopoli, A. Study of oils from Calabrian olive cultivars by chemometric methods. *Food Chem.* **2002**, *76*, 501–507.

(24) Kalua, C. M.; Allen, M. S.; Bedgood, D. R., Jr.; Bishop, A. G.; Prenzler, P. D. Discrimination of olive oils and fruits into cultivars and maturity stages on phenolic and volatile compounds. *J. Agric. Food Chem.* **2005**, *53*, 8054–8062.

(25) Gomez-Alonso, S.; Desamparados, S.; Fregapane, G. Phenolic compounds of cornicabra virgin olive oil. *J. Agric. Food Chem.* **2002**, *50*, 6812–6817.

(26) Nagy, K.; Bongiorno, D.; Avellone, G.; Agozzino, P.; Ceraulo, L.; Vekey, K. High performance liquid chromatography−mass spectrometry based chemometric characterization of olive oils. *J. Chromatogr., A* **2005**, *1078*, 90–97.

(27) European Communities, Regulation 510/2006. *Off. J. Eur. Communities* 2006, L 93/12.

(28) Vigli, G.; Philippidis, A.; Spyros, A.; Dais, P. Classification of edible oils by employing ^{31}P and ^{1}H NMR spectroscopy in combination with multivariate statistical analysis. A proposal for detection of seed oil adulteration in virgin olive oils. *J. Agric. Food Chem.* **2003**, *51*, 5715–5722.

(29) Christophoridou, S.; Dais, P. A novel approach for detection and quantification of phenolic compounds in olive oil based on ^{31}P NMR spectroscopy. *J. Agric. Food Chem.* **2006**, *54*, 656–664.

(30) European Communities, Regulation 2568/91. *Off. J. Eur. Communities* **1991**, *L 248*; **2003**, *L 1989*.

(31) Zwierzak, A. Cyclic organophosphorus compounds. I. Synthesis and infrared spectral studies of cyclic hydrogen phosphites and thiophosphites. *Can. J. Chem.* **1967**, *45*, 2501–2512.

(32) Montedoro, G.; Servili, M.; Baldioli, M.; Selvaggini, R.; Miniati, E. Simple and hydrolysable compounds in virgin olive oil. 1. Their extraction, separation and semiquantitative evaluation by HPLC. *J. Agric. Food Chem.* **1992**, *40*, 1571–1576.

(33) Johnson, R. A.; Wichern, D. M. *Applied Multivariate Statistical Analysis*; Prentice Hall: Englewood Clifss, NJ, 1998; pp 820.

(34) Kokkinofta, R.; Petrakis, P.; Mavromoustakos, T.; Theocharis, C. R. Authenticity of the traditional Cypriot spirit “Zivania” on the basis of metal content using a combination of coupled plasma spectroscopy and statistical analysis. *J. Agric. Food Chem.* **2003**, *51*, 6233–6239.

(35) Wilkinson, L. *Mobiles*; Department of Statistics, Northwestern University: Evanston, IL, 1985; pp 6.

(36) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984; pp 368.

(37) Fragaki, G.; Spyros, A.; Siragakis, G.; Salivaras, E.; Dais, P. Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using nuclear magnetic resonance spectroscopy and multivariate statistical analysis. *J. Agric. Food Chem.* **2005**, *53*, 2810–2816.

(38) Turland, N. J.; Chilton, L.; Press, J. R. *Flora of the Cretan Area: Annotated Checklist and Atlas*; The Natural History Museum: London, U.K., 1993; pp 439.

(39) Waterman, P. G.; Mole, S. Extrinsic factors influencing production of secondary metabolites in plants. In *Insect Plant Interactions*; Bernays, E. A., Ed.; CRC Press: Boca Raton, FL, 1995; pp 107−134.

(40) Minelli, S.; Maggini, F.; Gelati, M. T.; Angiolillo, A.; Cionini, P. G. The chromosome complement of *Olea europaea* L.: characterization by differential staining of eh chromatin and in-situ hybridization of highly repeated DNA sequences. *Chromosome Res.* **2000**, *8*, 615–619.

(41) Petrakis, P. V.; Touris, I.; Liouni, M.; Zervou, M.; Kokkinofta, R.; Theocharis, R.; Mavromoustakos, T. Authenticity of the traditional Cyprus spirit “Zivania” on the basis of ^{1}H NMR spectroscopy diagnostic parameters and statistical analysis. *J. Agric. Food Chem.* **2004**, *53*, 5293–5303.

(42) Banilas, G.; Moressis, A.; Nikoloudakis, N.; Hatzopoulos, P. Spatial and temporal expressions of two distinct oleate desaturases from olive (*Olea europaea* L.). *Plant Sci.* **2005**, *168*, 547–555.

(43) Okuley, J.; Lightner, K.; Feldman, K.; Yadav, N.; Lark, J. B. Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* **1994**, *6*, 147–158.

(44) Rubio de Casas, R.; Besnard, G.; Schönswetter, P.; Balaguer, L.; Vargas, V. Extensive gene flow blurs phylogeographic but not phylogenetic signal in *Olea europaea* L. *Theor. Appl. Genet.* **2006**, *113*, 575–583.

(45) Harmatha, J.; Nawrot, J. Insect feeding detrent activity of lignans and related phenylpropanoids with a methylenedioxyphenyl (piperonyl) structure moiety. *Entomol. Exp. Appl.* **2002**, *104*, 51–60.

(46) Brenes, M.; Hidalgo, F. J.; Garcia, A.; Rios, J. J.; Garcia, P.; Zamora, R.; Garrido, A. Pinoresinol and 1-acetoxypinoresinol, two new phenolic compounds identified in olive oil. *JAOCS* **2000**, *77*, 715–720.

Received for review October 8, 2007. Revised manuscript received February 14, 2008. Accepted February 15, 2008. We thank the Greek General Secretariat of Research and Technology and the European Union for financial support through Programs GKOR-017 to P.V.P. and 4.2.6.2B EPAN to P.D. Also, we are indebted to the program Regional Pole of Innovation for Value Creation funded by the administration of the Region of Crete and the European Union.

JF072957S