Detailed and exhaustive study of the authentication of European virgin olive oils by SEXIA expert system

By **R. Aparicio, V. Alonso and M.T. Morales**
Instituto de la Grasa. Avda. Padre García Tejero, 4. 41012 Seville. Spain.

RESUMEN

Un estudio detallado y exhaustivo de la autentificación de aceites de oliva virgen europeos mediante el sistema experto SEXIA

Se ha estudiado la autentificación de aceites de oliva virgen de diferentes regiones de España, Italia y Portugal, por su contenido en ácidos grasos, alcoholes, esteroles, metil esteroles e hidrocarburos. Se aplicaron métodos estadísticos multivariantes junto a la Teoría de la Evidencia. El estudio mostró una mejora en la capacidad predictiva utilizando esta teoría frente a otros métodos o sistemas expertos que no implementan la teoría de la posibilidad. Se ha realizado un estudio detallado y exhaustivo con aceites de oliva virgen italianos (Toscana y Basilicata), portugueses y españoles. Los resultados numéricos se muestran sobre mapas geográficos de las diferentes regiones estudiadas.

PALABRAS-CLAVE: Aceite de oliva—Autentificación —Estadística—Sistema experto SEXIA.

SUMMARY

Detailed and exhaustive study of the authentication of European virgin olive oils by SEXIA expert system

The authentication of extra virgin olive oils from different regions of Spain, Italy and Portugal, by means of their fatty acids, alcohols, sterols, methyl sterols and hydrocarbons content, has been investigated. Multivariate statistical methods and Evidence's Theory were applied. The comparative study shows greater predictive ability using this theory than the traditional statistical methods or expert systems that do not implement the possibility theory. A detailed and exhaustive study of Italian (Tuscany and Basilicata), Portuguese and Spanish virgin olive oils has been made. Geographically coloured maps of the studied regions are shown to strengthen the numerical results.

KEY-WORDS: Authentication—Olive oil —SEXIA expert systems — Statistics.

1. INTRODUCTION

Since ancient times, the virgin olive oil has been appreciated by the consumers, as much by its nutritional value as by its organoleptic characteristics. Its chemical composition is influenced by the climatology, altitude, soil composition etc. from origin zone. In consequence, there is an increasing interest in the geographical classification of olive oil as a reliable olive oil authentication may encourage bottling of good quality oils a way similar to "appellation d'origine" wines. In previous works (Derde et al., 1984; Tsimidou and Karakostas, 1993; Alonso and Aparicio, 1993) methyl esters of fatty acids were used for the chemical characterization of virgin olive oil. However, the information obtained from these chemical compounds is sometimes insufficient to characterize virgin olive oils from geographical zones too near one another, i.e. the different regions of a province.

In this work the authentication of extra virgin olive oils from different regions of Spain, Italy and Portugal, by means of their fatty acids, alcohols, sterols, methyl sterols and hydrocarbons content, was investigated. The results were analysed by using both multivariate statistical methods and an expert system which use the Evidence's Theory (Aparicio, 1988). The comparative study shows a greater predictive ability applying this theory than the traditional statistical methods applied in food characterization or expert systems that do not implement the possibility theory (Sabater et al., 1986; Derde et al., 1987).

2. EXPERIMENTAL

712 samples of virgin olive oil (EC., 1991), collected from Spain [426], Portugal [140] and Italy [146], have been characterized by up to fifty three chemical compounds described in Table I. All the chemical compounds were quantified using the procedures described in Aparicio and Alonso (1994).

3. THE SEXIA EXPERT SYSTEM

3.1 Introduction

The drawbacks of pure probabilistic methods led the researchers to consider new alternative approaches such as Evidence Theory or Fuzzy Logic. The SEXIA expert system applies the former theory with the aim of avoiding valid conclusions being drawn by chance. An extensive description of the structure of SEXIA expert system has been published previously (Aparicio, 1988; Aparicio and Alonso, 1994). So this present work describes its most important points so that the reader can understand how SEXIA runs.

Table I. **Chemical components of Virgin Olive Oil**

COMMON NAME	SYSTEMATIC NAME	EMPIRICAL FORMULA
PALMITIC	Hexadecanoic Acid	$C_{16}H_{32}O_2$
PALMITOLEIC	cis-9-hexadecenoic Acid	$C_{16}H_{30}O_2$
MARGARIC	Heptadecanoic Acid	$C_{17}H_{34}O_2$
MARGAROLEIC	cis-9-heptadecenoic Acid	$C_{17}H_{32}O_2$
STEARIC	Octadecanoic Acid	$C_{18}H_{36}O_2$
OLEIC	cis-9-octadecenoic Acid	$C_{18}H_{34}O_2$
LINOLEIC	cis-9-cis-12-octadecadienoic A.	$C_{18}H_{32}O_2$
LINOLENIC	cis-9-cis-12-cis-15-octadecatrienoic A.	$C_{18}H_{30}O_2$
ARACHIDIC	Eicosanoic Acid	$C_{20}H_{40}O_2$
GADOLEIC	cis-9-eicosenoic Acid	$C_{20}H_{38}O_2$
BEHENIC	Docosanoic Acid	$C_{22}H_{44}O_2$
DOCOSANOL	Docosanol Alcohol	$C_{22}H_{46}O$
TETRACOSANOL	Tetracosanol Alcohol	$C_{24}H_{50}O$
HEXACOSANOL	Hexacosanol Alcohol	$C_{26}H_{54}O$
OCTACOSANOL	Octacosanol Alcohol	$C_{28}H_{58}O$
TARAXEROL	D-triedoolean-14-en-3-ol Al.	$C_{30}H_{50}O$
DAMMARADIENOL	Dammara-20,24-dien-3-ol Al.	$C_{30}H_{49}O$
β-AMIRINE	Olean-12-ene-3β-ol Al.	$C_{30}H_{50}O$
BUTYROSPERMOL	5α, 13α, 14β, 17α, 20S-lanost-7,24-dien-3β-ol	$C_{30}H_{50}O$
24-METHYLENE-24-DIHYDRO-LANOSTEROL	Lanosta-8-en24-methylene-3-ol	$C_{31}H_{52}O$
CYCLOARTHENOL	9,19-cyclo-5α,9β-lanost-24-en-3β-ol Al.	$C_{30}H_{50}O$
24-METHYLENE-CYCLOARTHANOL	9,19-cyclo-24-methylene-5α,9β-lanostan-3β-ol Alcohol	$C_{31}H_{52}O$
CYCLOBRANOL	9,19-cyclo-24-ethyl-5α,9β-lanostan-3β-ol	$C_{31}H_{54}O$
CAMPESTEROL	24-methyl-colest-5-en-3β-ol Sterol	$C_{28}H_{48}O$
δ-5-AVENASTEROL	24-ethyliden-colest-5-en-3β-ol Sterol	$C_{29}H_{48}O$
β-SITOSTEROL	24-ethyl-colest-5-en-3β-ol Sterol	$C_{29}H_{50}O$
STIGMASTEROL	24-ethyl-colest-5,22-dien-3β-ol Sterol	$C_{29}H_{48}O$
OBTUSIFOLIOL	4α, 14α,dimethyl-24-methylene-5α-colest-8-en-3β-ol	$C_{30}H_{50}O$
GRAMISTEROL	4α-methyl-24-methylene-5α-colest-7-en-3β-ol	$C_{29}H_{48}O$
CYCLOEUCALENOL	9,19-cyclo-4α, 14α, dimethyl-24-methylene-5α, 9β-colestan-3β-ol	$C_{30}H_{50}O$
24-ETHYLLOPHENOL	4α-methyl-24-ethyl-5α-colest-7-en-3β-ol	$C_{30}H_{52}O$
CITROSTADIENOL	4α-methyl-24-ethyliden-5α-colest-7-en-3β-ol	$C_{30}H_{50}O$
A. OLEANOLIC	3β-hydroxiolean-12-ene-28-oic Acid	$C_{30}H_{48}O_3$
PHYTOL	3,7,11,15-tetramethyl-2-hexadecen-1-ol	$C_{20}H_{40}O$
ERYTHRODIOL	Olean-12-ene-3β,28-diol	$C_{30}H_{50}O_2$
COPAENE	Tricyclo[4.4.0.$0^{2,7}$]-dec-3-en-1,3-dimethyl-8-(1-methylethyl)	$C_{15}H_{24}$
VALENCENE	Naphthalene,1,2,3,5,6,7,8,8a -octahydro-1,8α-dimethyl-7-(1-methylethenil)-,[1R-(1α, 7β, 8aα)]	$C_{15}H_{24}$
MUUROLENE	Naphthalene, decahydro-1,6-bis(methylene)- 4-(1-methylethyl),[4R-(4α, 4aα, 8aα)]	$C_{15}H_{24}$
TRIDECENE	1-Tridecene	$C_{13}H_{26}$
HEPTADECENE	Heptadec-8-ene	$C_{17}H_{34}$
HENEICOSANE	Heneicosane	$C_{21}H_{44}$
TRICOSANE	Tricosane	$C_{23}H_{48}$
TETRACOSANE	Tetracosane	$C_{24}H_{50}$
PENTACOSANE	Pentacosane	$C_{25}H_{52}$
HEXACOSANE	Hexacosane	$C_{26}H_{54}$
HEPTACOSANE	Heptacosane	$C_{27}H_{56}$
OCTACOSANE	Octacosane	$C_{28}H_{58}$
NONACOSANE	Nonacosane	$C_{29}H_{60}$
TRIACONTANE	Triacontane	$C_{30}H_{62}$
HENTRIACONTANE	Hentriacontane	$C_{31}H_{64}$
DOTRIACONTANE	Dotriacontane	$C_{32}H_{66}$
TRITRIACONTANE	Tritriacontane	$C_{33}H_{68}$
PENTATRIACONTANE	Pentatriacontane	$C_{35}H_{72}$

3.2 Organization of the data

The taxonomic organization of the knowledge takes the form of a tree graph. Each node stores, in frame structures (Aparicio, 1988; Aparicio and Alonso, 1994), the information about the class finding.

A frame is an abstract specification for a class. Each entity of a frame consists of a name, its attributes and the values linked with these attributes. The domain-specific components store the substantive characteristics of each parameter. Let us display the node corresponding to the tetracosanol parameter:

superclass: aliphatic_alcohols
class: tetracosanol
[DOMAIN-SPECIFIC COMPONENTS]
(universe_of_discourse Tuscany)
(numeric_range (boundaries - and - peakpoint
14.440 36.850 27.412))

The universe of discourse slot is related to Tuscany (Italy) and tetracosanol. The numeric_range slot displays summarized information on the chemical compound through the peak-point and bandwidths-left & right-of its probability distribution.

3.3 Representation of the rules

The SEXIA knowledge base is represented by rules obtained from the database frame. A rule expresses a conditional relationship between the propositions (premises) and the conclusion. In other words every rule is represented by means of a set of propositions which, if they are fulfilled, provide a conclusion associated with a value (CF means certainty factor) that indicates its certainty. The external representation of a rule is:

[(RULE <rule-name>)
(IF <propositions>)
(THEN <conclusions>)
(CF = <value>)]

Obviously each rule needs to verify its propositions to reach a conclusion that, in our case, is linked with the olive oil identification process.

The first part of Figure 1 shows the general finding of SEXIA concerning the frame knowledge base. There are three types of information: (i) varieties of olive tree; (ii) localities of olive growing zones; and (iii) general information collected from the bibliography (M.A.P.A., 1988; Morettini, 1950; Barranco and Rallo, 1984; Cadastro oleicola, 1983).

The second part of Figure 1, concerning Spain, shows how level 1 (zones) is reached from level 4 (countries). At each one of these levels the expert system has rules to discern the best node to be selected at the successive levels. The rules are structured in four types: inexact reasoning, relational, lineal and heuristic. A detailed description appears in Aparicio and Alonso (1994).

This work displays the expert system knowledge base of some of the nodes displayed in figure 1: (i) Italy (the provinces of Tuscany and Basilicata); (ii) Portugal (all the country); and (iii) Spain (its Autonomous Communities and the Andalusian province of Jaén). Thus, we can observe

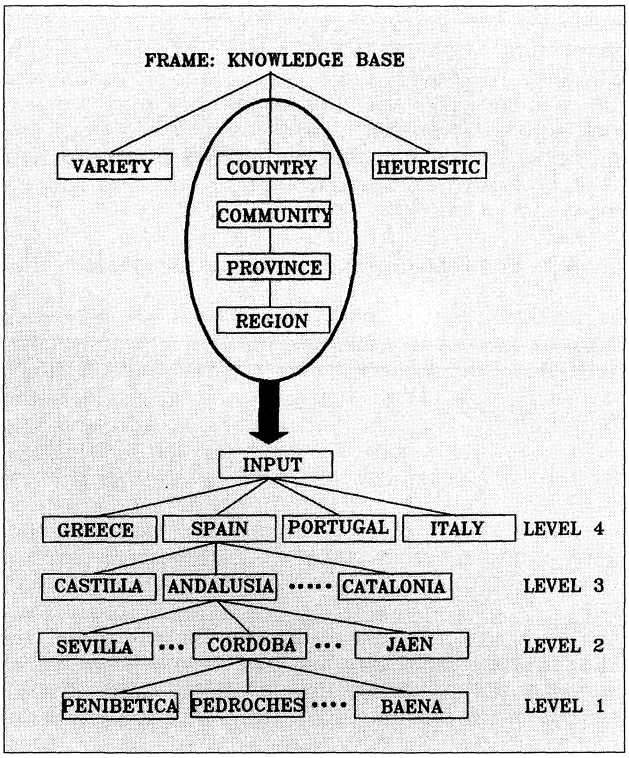

Figure 1
Tree graph of the knowledge base

the ability of the expert system in characterizing olive oils from countries to olive grove zones. The knowledge base of each node also allows an unknown sample to be identified as belonging to a node (country, province, zone, etc.) when expert reaches this node by a searching process (Aparicio, 1988; Aparicio and Alonso, 1994).

4. ALGORITHMS USED IN FOOD CHARACTERIZATION/AUTHENTICATION

The objective of this section is to analyse the competence of the most commonly used statistical procedures versus Evidence Theory characterizing the Tuscan olive oil. The latter theory only uses, in this section, the linear equations of the knowledge base stored at the Tuscany node of SEXIA expert system.

The selected statistical procedures were: Stepwise Linear Discriminant Analysis (SLDA), Factor Analysis (Principal Components) and Cluster Analysis, which are the most commonly used in foodstuff characterizations (Bisquerra, 1989; BMDP, 1981).

4.1 Principal Components Analysis

Principal Components Analysis (PCA) is the most commonly applied procedure in characterization. In this work, we previously checked by the Barlett's Sphericity and Kevin-Meyer-Olkin tests that PCA might be applied to the data set. PCA was applied under the following conditions: Kaiser's normalization, varimax rotation, tolerance limits for matrix inversion (0.0001) and cross-

validation. With these conditions, the R-type extracted five significant eigenvalues that explained 80.4% of total variance. However, there is no clear differentiation between the eight Tuscan provinces, due to each one of the eigenvectors partially explains more than one province. A posterior cluster analysis on the results of Q-type, now using eight eigenvectors, did not also add any significant information.

4.2 Stepwise Linear Discriminant Analysis

This statistical procedure allows the levels of correct classification of the samples in their provinces to be computed. We have applied the most rigorous conditions in order to avoid the possibility of the results being obtained by chance. The criterion used was the F_to_enter (3.29) value, obtained from the F-distribution for the number of cluster (m=7) and samples per province (n=9) for a probability greater than 97.5%.

The values of correct classifications in prediction, after applying the Jackknife algorithm, were: Arezzo (66.7%), Firenze (30.0%), Grosseto (46.7%), Livorno (93.8%), Lucca (90.0%), Pisa (66.7%), Pistoia (64.3%) and Siena (86.7%), the total being 70.8%. These rather low classification values arise because the procedure has to formulate a mathematical equation which classifies each sample into its province. Thus, the greater the number of groups the lower the percentage of correct classifications. Moreover, if some chemical parameters change over the years, then the equation might be rendered invalid for characterizing samples.

4.3 Evidence Theory

Comparison of the different theories requires similar conditions and so the decision equations for this work were obtained from SLDA (BMDP, 1981), using the same conditions as those applied to the SLDA study, but now comparing two-by-two all Tuscan provinces, e.g. Arezzo-Firenze, Arezzo-Lucca and so on. Hence, the expert system explained in this paper works with bivariate distributions. We have used this type of comparison for two reasons: (i) the maximum level of differentiation is obtained from two-by-two comparisons; and (ii) the number of two-by-two comparisons is lower than, or at best equal to, other possible combinations (C(8,2)=28, C(8,3)=56, etc.).

In the case of Tuscany, the expert system uses twenty eight canonical rules. Any sample has, therefore, seven associated values of probability of being classified in each one of the eight Tuscan provinces (Figure 2). Thus, besides the strict conditions with which each one of the SLDA probabilities was computed, the system verifies the correct classification of a sample in each one of the provinces seven times. Later, Evidence Theory works with the whole set of information to compute the belief intervals of all samples in each one of the provinces. Henceforth, the final certainty will be displayed in terms of probability.

Despite the rigorousness of this new procedure its results are better than statistical procedures: Arezzo 93.3%, Firenze 80.0%, Grosseto 86.7%, Livorno 93.8%, Lucca 85.0%, Pisa 86.7%, Pistoia 92.9%, Siena 95% (in fact 99.9%) and the total probability 90.0%. Moreover,

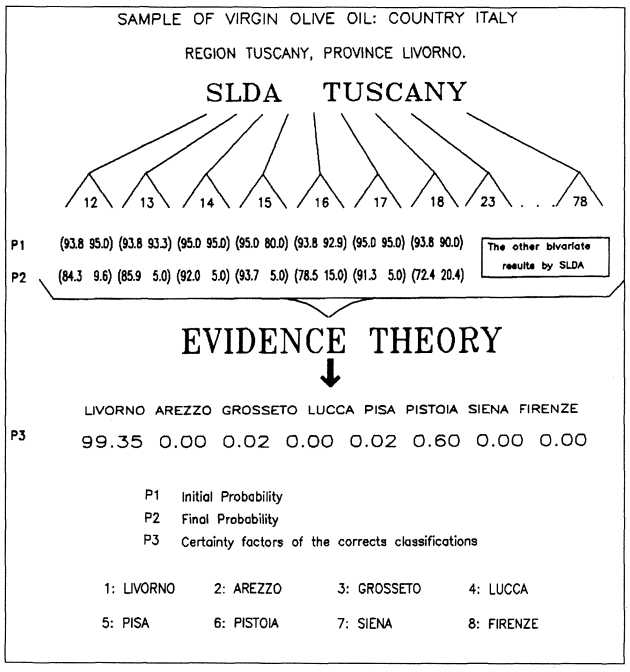

Figure 2
A piece of the tree structure of SEXIA Expert System

Evidence Theory shows some advantages: (i) the verification is more strict than the statistical procedures; (ii) the results are better protected against the changes in the chemical compounds over the years; (iii) the probability of correct classifications does not depend on the number of categories to be characterized; and (iv) if a new category is added to the set it is not necessary to correct the other equations, it being sufficient to add the new equations that discriminate this category from the others.

5. RESULTS

Analysis and representation of the results of discriminant procedure

The number of categories analysed would involve an excessive number of tables if all the results were described. For this reason the results of Tuscany (Italy) will be used to explain the process. Similar tables were built from the other studies. For example, Table II shows the levels of significance, in linguistic terms, with which it is possible to assure that a chemical parameter can characterize a province of Tuscany. Table III shows the percentage of samples correctly classified by prediction in their provinces, using a bivariate analysis. In order to avoid good classifications by chance, the values of F_to_Enter and F_to_Remove (values of F distribution) were chosen according to the number of samples from the province which had the lower number of samples of the pair. CAD software (López and Tajadura, 1992) was used to plot the results on the geographical maps. This work studies levels 1 and 2 corresponding to the "provinces" and "zones" categories (Fig.1).

Table II. **Discriminatory capacity of chemical parameters in two-by-two comparison of the provinces of Tuscany**

Provinces of Tuscany	1 2	1 3	1 4	1 5	1 6	1 7	1 8	2 3	2 4	2 5	2 6	2 7	2 8	3 4	3 5	3 6	3 7	3 8	4 5	4 6	4 7	4 8	5 6	5 7	5 8	6 7	6 8	7 8
PALMITIC	4	5	5		5	3	5		5	3				5	4				5	5	5	5	4	1	5			
PALMITOLEIC		5	5					5	4		2			1	5		5	5	5		5	5	5			3	3	
STEARIC				5			5			5			5		1			1	2	2		4	5	4		1	5	5
OLEIC	2	5	5	1	5	1	5	5	5		2		5	5	5		3		5	5	5	5	2		5	1		2
LINOLEIC		4	5					4	5				1	4	5	2	5		5	5	5	5			5			4
LINOLENIC			5	5		4			4	4		2	1	5	5	1	5	5										
ARACHIDIC				1	1					5	4		3		1					2		1		2		1		
GADOLEIC										1					3	1												
DOCOSANOL	1	1						5	5	5	4	1						4			1	1			1			
TETRACOSANOL	1		5				1	5	5	4	3		5	3			5		5	5	5	4		1				3
HEXACOSANOL		5	5	1	2		1	5	5	5	5		5		4	1	5	3	1		5			1		3		1
OCTACOSANOL									2					1					5	5	5	5						
CYCLOARTHENOL				5	1	1				5	1	1		4	5				5	4		5	1	1	5			
24METHYLENCYC				5						5		1		1	5				5		4	4	5		5	1		
CAMPESTEROL			2				5		4		2	1	2					5	5			5	4	2	5		5	5
STIGMASTEROL		1	3	4	4	4																						
βSITOSTEROL		5			2		5	5		2	2		5	5		2	5				1	2		4		4		5
δ5AVENASTEROL		5						5	1	1	2		2		3	2	5				2			2				3
ERYTHRODIOL		1	1	5	1	5				5		2			5				5		1		5	5	5			
CITROSTADIENOL																						1						
CYCLOBRANOL		2	2		5					1	5					4	3		1	5	5		4	4	1	5	5	

Linguistic terminology	Significance level	Sign of the table
completely significant	p <= 0.001	5
very significant	0.001 < p < 0.005	4
significant	0.005 < p < 0.01	3
slightly significant	0.010 < p < 0.02	2
very slightly significant	0.02 < p <= 0.05	1

5.1. Italian Olive Oils

Comparison between Italian regions using fatty acid has been reported (Alonso and Aparicio, 1993). This paper describes the authentication of virgin olive oils from provinces of Tuscany and Basilicata regions, using the compounds described above (Table I).

Table III. **Results of the classifications achieved after applying SLDA to provinces of Tuscany**

TUSCANY	Firenze	Grosseto	Livorno	Lucca	Pisa	Pistoia	Siena
Arezzo	80.00	99.00	96.77	94.29	99.00	99.00	99.00
Firenze		99.00	96.15	99.00	96.00	91.67	99.00
Grosseto			93.55	99.00	86.67	96.55	99.00
Livorna				99.00	93.55	96.67	99.00
Lucca					99.90	94.12	97.14
Pisa						96.55	99.00
Pistoia							99.00

5.1.1. Characterization of Tuscany provinces

120 samples were collected from eight provinces of Tuscany and characterized by 21 chemical compounds. Table II shows the level of significance of each one of these chemical compounds distinguishing the provinces of Tuscany two-by-two: 1=Arezzo, 2=Firenze, 3=Grosseto, 4=Livorno, 5=Lucca, 6=Pisa, 7=Pistoia, 8=Siena. Therefore, it is possible to know the value of each chemical compound in discriminating among the provinces of Tuscany, for example:

-Stigmasterol sterol discriminates olive oils produced in Arezzo from those produced in other provinces.

-Docosanol and Tetracosanol allow olive oils produced in Firenze to be distinguished from those produced in Livorno, Grosseto and Lucca.

-The lowest values of oleic acid and octacosanol alcohol belong to the olive oil harvested in Livorno.

-Erythrodiol characterizes the olive oil from Lucca, which has the lowest values for this compound of all the provinces of Tuscany.

-Cyclobranol may characterize the olive oil harvested in the province of Pisa.
-etcetera.

Relational rules (Aparicio et al., 1981), such as those described in Table IV, were attained using the values of chemical parameters. The relational rules were also used by the expert system to characterize the Tuscan olive oils.

Table IV. **Some of decision rules built with the data from Italian oils**

	TUSCAN DECISION RULES		the sample is from		probability
IF	Oleic/Cyclobranol	< 10	THEN	Pisa	71.4%
IF	Palmitic/Log(Octacosanol+1.0)	> 5	THEN	Livorno	78.6%
IF	βsitosterol/Cyclobranol	< 12	THEN	Pisa	71.4%
IF	Palmitoleic/Gadoleic	>3.5	THEN	Grosseto	72.7%
	BASILICATA DECISION RULES	**the sample is from**			**probability**
IF	Stigmasterol	> 2	THEN	Potenza	75.0%
IF	Erythrodiol	> 3	THEN	Potenza	76.9%
IF	Phytol	> 10	THEN	Matera	83.3%
IF	Linoleic	> 8	THEN	Matera	87.5%

Figure 3 shows the results of Evidence Theory. A different colour was assigned to each province. Thus, if one sample is classified in two provinces with certainties "x" and "y", its colour is a mixture of the colours assigned to each of the two provinces and the percentage of each colour is determined by the values of x and y.

The province of Grosseto was given the colour yellow. After applying Evidence Theory to the Tuscany data 86.6% of the Grosseto samples were in terms of probability correctly classified. They are thus represented in yellow on the final map.

This province is crossed by various rivers (Ombrone, Alberga, Bruna, Fiora, etc.) that form valleys. The soil composition in the valleys, which is different to that in the mountains, will affect the chemical composition of the olive oil (Ferreiro and Aparicio, 1992; Aparicio et al., 1993). Thus, the two samples with a pink tone, that belong to the most mountainous zones (close to Mount Amiata of approx. 1700m and Monte d'Alma of approx. 550m), were classified in Grosseto but with a low probability (P=34%, P=44%), whereas the rest of the samples, from the valley zones, were classified in Grosseto with probabilities higher than 85%. One possible chemical explanation for this could be the lower Cyclobranol content (<1.0) in the oil from the higher altitude zone, this being closer to the mean content of the oil from the mountainous zone of Arezzo (1.109) than to that of the oils from Grosseto (4.894).

The contiguous region on the map, with a fuchsia colour, corresponds to the province of Siena. Almost all the samples of this province were correctly classified (P>86%).

Further to the North is the province of Arezzo, the initial colour of which was red. 14 out of 15 samples from this province were correctly classified,the final certainty being greater than 90% for the Northern samples though in the Southern zone (Val di Chiana) this was lower (P>72%).

The initial colour assigned to Firenze was violet. As can be seen, there is a zone with this colour but in various tones. This indicates that not all of the samples were classified as being certainly from this province, only four from the valley of the river Arno have value greater than P>90%. Thus, the final colours obtained are a mixture of those assigned to other provinces.

The figure shows a narrow band of green. These are the samples from Livorno, 93.7% of which were correctly classified and are thus represented in the figure with the colour initially assigned to this province. Only one sample appears with a different colour, this being blue, due to it is classified with 89% certainty in Pistoia.

Sky-blue was assigned to the province of Pisa. 87% of the samples from this province were correctly classified and appear in the figure with this colour. The samples appearing in yellow are those classified in Grosseto with a certainty of more than 90%.

To the north of Tuscany is the province of Pistoia. 71.4% of the samples from this province were correctly classified with a certainty of more than 90%. The initial colour of samples from this province was blue. Despite the high number of correct classifications a certain number of samples were incorrectly classified, due to high levels of methylene-cyclo-arthanol or, low values of oleic acid and tetracosanol.

The grey zone in the figure is situated at the limits of the province of Lucca. 80% of the samples from this province were correctly classified and are thus represented by the colour initially assigned to Lucca. The samples of oil from this province were largely collected from low altitude (valley) zones, influenced by the proximity of the mountain chains to the north (Apennines Tosco Emiliano). These zones produce oils with a characteristic chemical composition which differentiates them from the other provinces. Thus, the levels of erythrodiol are <3.0 in all of the samples analysed in this province, and those of methylene-cyclo-arthanol are >60.0 in 85% of the samples (Ferreiro and Aparicio, 1992).

5.1.2. Characterization of olive oils from Basilicata

Basilicata is a Southern Community of Italy with two provinces: Potenza and Mattera. 13 samples from Potenza and the same number from Mattera were characterized using the following series: fatty acids, alcohols, sterols and methyl sterols.

The chemical compounds linoleic acid, phytol and erythrodiol alcohols and stigmasterol sterol proved their value in characterizing the provinces, either by linear equations (the use of linoleic acid and erythrodiol resulted in 88.46% of samples being correctly classified by prediction), or by relational decision rules (Table IV).

5.2 Portuguese olive oils

140 samples were collected from the 18 Portuguese provinces. Due to the large number of provinces and the few samples from some of them, we grouped the samples into six regions according to geographical criteria. Thus, we studied these clusters by PCA before defining these as the initial ones for the SEXIA Expert System.

Before using PCA, the values of the Kaiser-Meyer-Olkin test (0.72) assured that PCA could be used. Three factors were computed after cross-validation, explaining 81% of total variance. Figure 4 shows 3D plots of the six regions independently (the axes are the three principal components).

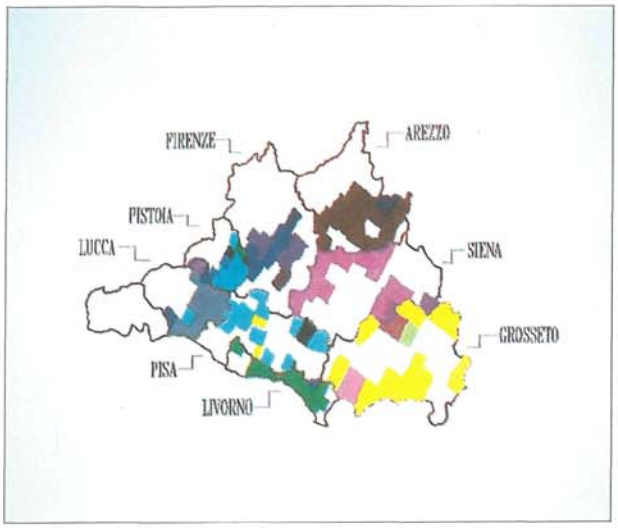

Figure 3
Map of Tuscany (Italy)

If all results were displayed in one plot, then we would not be able to distinguish the regions except for Litoral. However, the 3D plots can be useful for characterizing the Portuguese olive oils and for seeing how the expert systems can improve the pure probabilistic methods.

Analysing Fig. 4, it can be seen that the Alentejo-Algarve plot, which represents the Southern Portuguese olive oils, shows two great peaks: (i) one on axis +zz', produced by the samples from Algarve; and (ii) the other on axis -zz', produced by samples from Alentejo. The small peak on axis +zz' represents the samples from the coastal province of Setubal.

The province of Ribatejo is characterized by two peaks and hence two different types of olive oil. The peak on the +zz' semi-axis represents the samples from Santarem (close to the coastal province of Litoral) whilst the inland samples are represented by the peak on -zz' axis.

The olive oil harvested in the provinces of Litoral and Estremadura seem to be very homogenous. Only one peak on the +zz' semi-axis characterizes these provinces.

The plot of Beira Baxa shows two peaks. The peak on the +zz' axis belongs to the samples close to Leiria, the province of Litoral. The other samples from the inland province of Beira Baxa generate the peak on the -zz' axis. The former samples were collected from the mountains (Sierra de la Estrella), whilst the latter from the valleys.

The 3D plot of Beira Alta only shows peaks on the -zz' semi-axis. One of them represents the samples from the Northern provinces of Viseu and Guarda. The other peaks represent the samples from Eira and South Beira Alta.

The provinces of Minho, Duoro and Tras os Monte have been grouped into one region (North of Portugal). However, the 3D plot displays the differences between the olive oils from the three groups. The peaks on the +zz' semi-axis belong to the coastal provinces. The highest peak on semi-axis +zz' is due to the samples from Minho,

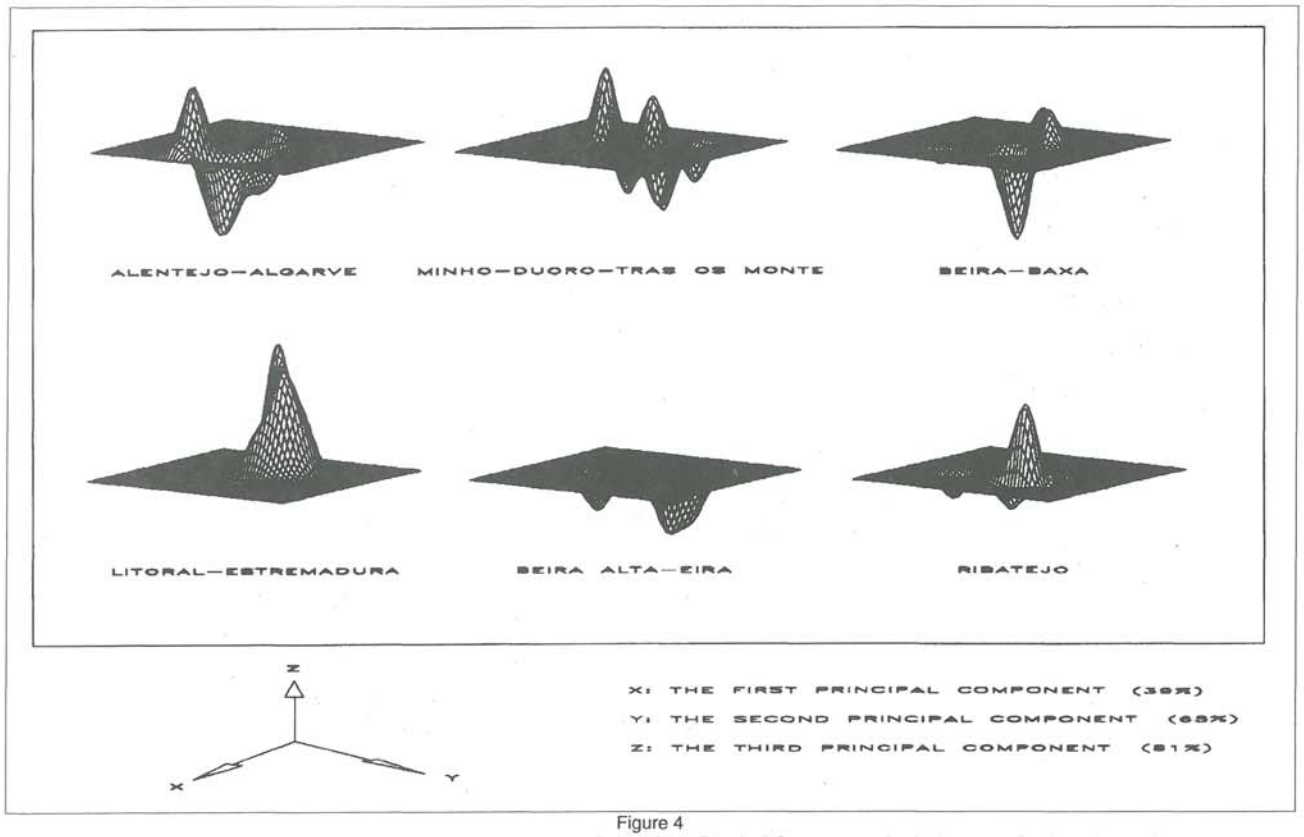

Figure 4
Representation of the first three eigenvectors achieved after applying Principal Components Analysis to samples from Portugal.

The central coastal region of Portugal is very different from the other areas. On the other hand, Beira Baxa, situated to the right, does not appear on the map with a very precisely defined colour.

There are two zones within Beira Alta. One of these to the North, close to Tras os Monte, appears in dark pink and the other to the South is shown in pale pink. In the latter such parameters as altitude, proximity to the coast, climate etc. (Ferreiro and Aparicio, 1992; Aparicio et al., 1993) mean that the samples from this region are classified both in Beira Baxa and in Litoral.

Finally, as has been explained previously for the results of principal components analysis, the three northern provinces considered as one group are really quite different despite their similar latitudes. That region coloured in red on the map is made up of the samples from Tras os Monte, which is a fairly homogenous unit extending to some parts of the north of Beira Alta. Duoro, on the other hand, is classified as being in Litoral-Estremadura and the colour given to this region is yellow. Being close to the coast, a low altitude zone, the chemical composition of the oils from this region has no similarity to those from Tras os Monte. Minho, to the North-West, is not classified in any one province more than another, which explains its colour on the map.

In conclusion, the percentage of correct classifications obtained by the Expert System exceeds those obtained using PCA.

5.3. Spanish olive oils

Spain is the principal world producer of olive oil with an average production of 500,000 tons a year. Andalusian olive oil represents 80% of this production, i.e. 33% of the total European Community production. Thus, we have divided this study into two groups, the non-Andalusian provinces and the Andalusian provinces.

5.3.1. Characterization of non-Andalusian provinces

Eighty seven samples were collected from eight provinces of Spain: the midland provinces of Cáceres, Badajoz, Toledo and Ciudad Real, and the North-eastern provinces of Castellón, Tarragona, Zaragoza and Teruel. These two groups have quite different climates, soils and altitudes.

An analysis of variance of the chemical compounds quantifying the dataset has allowed us to conclude that:

-the fatty acids oleic and linoleic characterize the midland provinces of Toledo and Ciudad Real;

-the olive oils produced in the province of Toledo are characterized by the methyl sterols, Cycloeucalenol, Citrostadienol and 24-Ethyllophenol;

-the fatty acid arachidic characterizes the olive oils from Teruel;

-the triterpenic alcohol 24-Methylene-24-dihydro-lanosterol characterizes olive oils from Castellón; and

-the olive oils from the northern coastal provinces of Tarragona and Castellón are characterized by the triterpenic alcohol Taraxerol.

Relational rules were built. They are shown in Table V.

The prediction probability in the differentiation of the eight provinces two-by-two are very high, except for the pair Teruel-Zaragoza with 82%. Thus, the certainty associated with the final discrimination obtained by Evidence Theory is high in every case.

Table V. **Some of the relational rules built with the data from Spain**

	SPANISH DECISION RULES		**the sample is from**	**probability**
IF	Linoleic/Campesterol > 3	THEN	Cáceres	71.4%
IF	10*Margaroleic/Linoleic < 1 AND Stigmaterol/Taraxerol < 1	THEN	Cáceres	75.0%
IF	Stearic/Gadoleic < 4	THEN	Teruel	75.0%
IF	100*Margaroleic/Erythrodiol > 5.0 AND Stearic/Gadoleic < 4	THEN	Teruel	87.7%
IF	Linoleic/Obtusifoliol < 4 AND 10*Gramisterol/Cycloeucalenol < 2	THEN	Toledo	83.0%
IF	Cycloeucalenol/Citrostadienol > 1	THEN	Toledo	99.0%
IF	Cycloeucalenol/Octacosanol > 2	THEN	Toledo	77.0%
	ANDALUSIAN DECISION RULES		**the sample is from**	**probability**
IF	Phytol < Butyrospermol AND Phytol < Docosanol	THEN	Jaén	90.4%
IF	Hexacosanol > δ5Avenasterol	THEN	Jaén	77.6%
IF	Erythrodiol*10/Hexacosanol < 2.5	THEN	Jaén	83.9%
IF	Hexacosanol/Stigmasterol > 5 AND Linoleic/Linolenic < 10	THEN	Jaén	90.7%
IF	Erythrodiol*10/Tetracosanol < 5 AND Linoleic/Linolenic < 10	THEN	Jaén	85.9%
IF	Linoleic*100/βsitosterol > 5 AND Obtusifoliol*10/Erythrodiol < 5	THEN	Huelva	62.5%
IF	Palmitoleic*10/Campesterol < 1 AND Palmitoleic*10/Butyrospermol < 1	THEN	Málaga	60.0%
	JAÉN DECISION RULES		**the sample is from**	**probability**
IF	Palmitoleic/Tritriacontane > 5 AND Citrostadienol*10/Hexacosanol <5	THEN	La Campiña	70.0%
IF	Gramisterol*10/δ5Avenasterol > 1 AND Phytol/Stigmasterol < 2	THEN	La Campiña	66.6%
IF	βamirine/Tridecene > 7 AND Obtusifoliol/Tridecene > 4 AND Hexacosanol/Butyrospermol < 2	THEN	S. del Segura	62.5%

Figure 6 shows the results of applying Evidence Theory to the samples of these provinces. The colours initially assigned to each province were: blue (Badajoz), sky-blue (Cáceres), red (Castellón), yellow (Cuidad Real), fuchsia (Tarragona), violet (Teruel), green (Toledo) and grey (Zaragoza). Any of the samples studied could have been classified in any of these eight provinces. However, all of the samples were classified with a certainty greater than 90% in the province from which they were collected. In the figure each coloured territory indicates the area from which the samples were collected. Because of the results of the final classification, therefore, all of the samples from one province are represented by the colour initially assigned to that province.

One northern part of Castellón is shown in pink as the samples from this area were classified with a high certainty in Tarragona. The chemical composition of the oils from this area is different to that of the other samples from Castellón and is more similar to those from the contiguous province. Thus, the stearic acid and phytol values in the oils from this area are higher than those in the other oils from the same province but are similar to those in oils from Tarragona. 86% of the samples from Teruel were correctly classified, but there were some exceptions which were in large part classified in Zaragoza and are, therefore, shown on the map in a paler colour. These wrongly classified samples were characterized by having higher levels of docosanol and cycloarthenol than the other samples from Teruel, but similar levels to the mean values found in the samples from Zaragoza. This was also the case for the levels of phytol and palmitic acid.

5.3.2. Characterization of Andalusian provinces

339 samples collected from the Andalusian provinces of Córdoba, Jaén, Málaga, Huelva and Seville have been used in this characterization. The greatest number of samples were from Córdoba and Jaén, since these produce the majority of the world's olive oil.

With respect to the level of significance of some chemical compounds, analysis of variance has allowed those characterizing the Andalusian provinces to be identified:

-low values of aliphatic alcohols and high values of β-sitosterol characterize the Sevillian olive oils;

-the olive oils from Huelva (basically the Verdial variety) are easily characterized by phytol and erythrodiol, although there are other chemical compounds with a reasonable degree of significance such as gramisterol, cycloeucalenol, citrostadienol, β-amirine, butyrospermol, 24-methylene-cycloarthanol and 24-methylene-24-dihydro-lanosterol;

-ethyllophenol, citrostadienol, margaric and magaroleic acids in the samples from Jaén are low, and oleic is high. All of them are very significant in comparison with other provinces;

-Málaga province has low values of palmitic and palmitoleic. These compounds are completely significant in all cases (Aparicio et al., 1981).

-Córdoba has not got any chemical parameter very different from the others. This fact will have an effect in its later discrimination.

Some of the relational rules built with the data from Andalusia were shown in Table V.

The right-hand portion of Figure 6 shows those Spanish provinces belonging to the Autonomous Community of Andalusia, data from which has been considered in the present study.

The initial colours were fuchsia for Córdoba, blue for Jaén, yellow for Málaga, green for Seville and red for Huelva. Despite the high total number of samples [339], the results after applying Evidence Theory were very good, as can be seen in the figure. 91% of samples from Seville, 92% from Málaga, 94% from Jaén, 86% from Huelva and 64% of samples from Córdoba were correctly classified.

Córdoba was the province with the lowest number of correctly classified samples. Some of the samples from zones of Córdoba near to Jaén were classified partially or totally in the latter province (Aparicio et al., 1991). These samples therefore appear on the map in blue. This is possibly due to the fact that in Córdoba different varieties of olive are cultivated (Barranco and Rallo, 1984), each variety giving oils with a quite different chemical composition. This is not the case in Jaén, where the variety Picual is cultivated in almost the whole province. Despite the high number of samples analysed from Jaén, 86% of them were classified in their province of origin with a high certainty (greater than 90%).

One sample from Huelva (shown in brown in the figure), taken from a region close to Seville, was classified in both Huelva and Seville. As has been mentioned previously, the oils from Seville are characterized by their aliphatic alcohol content. In the case of the samples mentioned above, the hexacosanol value (3.45) is closer to that of oils from Seville (4.00) than of oils from Huelva (5.77).

One sample from the region of Málaga bordering Seville was classified in Seville with a final certainty of 85%. This is shown in the figure in green. A possible explanation for this may be that the altitude of the zone from which the sample was taken is different to that of the regions of Málaga (Ferreiro and Aparicio, 1992). As a result the methylene-cycloarthanol and cycloeucalenol content of the sample is lower than the mean value for the other samples from Málaga and similar to that of samples from Seville.

One sample from Seville, shown in pinky tone in the figure, was classified with a certainty of 60% in Córdoba and in Seville with a certainty of 40%. In this case the methylene-cycloarthanol value was high, and the cycloeucalenol value low, in comparison with other samples from Seville but not in comparison with samples from Córdoba.

5.3.3. Characterization of regions from Jaén province

Jaén is divided into eight different olive oil producing zones: La Campiña, La Loma, El Condado, Martos, Sierra de Cazorla-Quesada, Sierra Sur, Sierra Morena and Sierra del Segura. Table V displays some of the relational rules built with these data. Figure 7 shows the results of Evidence Theory.

The initial colours assigned to each zone were: green for La Campiña, sky-blue for Martos, grey for Sierra Sur, blue for Sierra Morena, yellow for La Loma, pink for El Condado, purple for Sierra de Cazorla and red for Sierra del Segura.

Eight zones can be clearly distinguished in Figure 7. To the left there is a green zone corresponding to the samples from La Campiña, 93% of which were correctly classified. Below this zone is a sky-blue zone consisting of the samples from Martos, 84% of which were correctly classified. Between the two aforementioned regions there is a blue-green zone which corresponds to a sample originating in Martos but which was classified with only a 56% certainty in that zone and with a certainty of 46% in La Campiña. This double classification explains the final colour used to denote it in the figure. In addition, another sample, also from Martos, was classified with a 67% certainty in La Loma and with a 31% certainty in Martos. This sample is represented by a yellow-green colour. Despite the proximity of La Campiña and Martos, they produce olive oils with chemical parameters which are quantitatively enough different. Among other factors the altitude plays a definite role in deciding the chemical composition of virgin olive oil (Ferreiro and Aparicio, 1992). La Campiña has a mean altitude of 346m, while that of Martos is almost double this at 641m. The blue-green sample originates in a zone at 441m, an altitude intermediate between the means of each of the two areas. The value of palmitoleic acid of this sample (0.85) is greater than the mean value for the same parameter in samples from Martos (0.76) and similar to that of samples from La Campiña (0.83). In the zone represented in yellow-green the oil have a high value for Gramisterol and Cycloeucalenol (0.94 and 5.33, respectively) in comparison with other samples from Martos and more similar to the values found in oils from La Loma. This fact will influence in its final classification.

To the South of the province the zone represented in grey constitutes 90% of the samples collected from the Sierra Sur. This area is mountainous with a mean altitude of 833m. One sample from this zone classified in Martos, was collected from an area with a mean altitude of 607m, lower than that of the Sierra Sur and more like that of Martos. From the chemical point of view, this oil has a value for linoleic acid lower than that of other samples from Sierra Sur and closer to those found in samples from Martos.

A central zone coloured in yellow can also be distinguished in Figure 7, this containing 70% of the samples from La Loma. Almost in the centre of the map, in the yellow zone, there is an area represented by a greenish colour. This is due to a sample from La Loma which was clasified with more than 50% certainty in Martos. The sample was collected from an area at an altitude of 769m, similar to the altitudes found in the north of Martos. Highly correlated with the altitude is the methylene-cycloarthanol value (Ferreiro and Aparicio, 1992). In the case of this sample the value for this compound (76.91) is low in comparison with the mean value for samples from La Loma (105.61) but not in comparison with the samples from Martos.

Below La Loma is a violet region in which 87% of the samples from Sierra de Cazorla are grouped. In a zone to the north of this region the predominant colour is greenish, which is due to a sample classified with a 45% certainty in La Loma and with a 35% certainty in Martos. The values of methylene-cycloarthanol in this sample (108.46) are similar to the values found in samples from La Loma and lower than the mean value for samples from Cazorla (125.06).

In the North, going from left to right, there is a blue zone that corresponds to all of the samples from the Sierra Morena. To the right a fuchsia coloured zone arises as a result of 75% of the samples from El Condado.

Finally, the red colour is produced by 83% of the samples from the Sierra del Segura. There were some samples from this region that were badly classified, such as one from a location in the south of the Sierra, which was classified in La Loma with 78% certainty. For this reason it is represented in yellow. Chemically this sample has hexacosanol and methylenlanosterol contents higher than other samples from the Sierra del Segura and similar to values found in the samples from La Loma. Another sample, this time from the north of the Sierra, was classified in El Condado and is, therefore, represented in fuchsia. This sample was from a location situated at 555m of altitude which is lower than other areas in the same region, the mean altitude of which is 759m, and similar to the mean altitude of El Condado (617m). Analysis of the chemical components of this sample showed that it had high values of cycloeucalenol (6.92). The mean values for this compound in samples from the Sierra del Segura and from El Condado are 4.88 and 5.52, respectively. The value for this compound in this badly classified sample was, therefore, closer to the mean for El Condado. On the other hand the cycloarthenol value for this sample (11.27) was low compared to the mean for this compound in samples from the Sierra del Segura (16.36) and more like that of samples from El Condado (12.08).

FINAL CONCLUSIONS

This paper has tried to explain how SEXIA expert system currently runs on a DEC-Station 5200 or μVAX II under Ultrix Operating System and Lisp language. Basically, SEXIA has allowed us to classify unknown samples and characterize different olive oils, diminish the effect of data dispersion over the years and give more information than the statistical programs.

On the other hand, real oil maps were shown. Initially, the zones were taken from geopolitic maps, since they are the only information reported. However, SEXIA allowed to cluster the groups according to the similarites or dissimilarities among samples and, in consequence, rebuilt the maps to attain true olive grown maps. SEXIA has also allowed to know the oils from the countries, regions or provinces, taken into account the varieties of their cultivars, climate and the edaphologics characteristics.

ACKNOWLEDGEMENTS

The authors would like to acknowledge their indebtedness to all those who have collaborated in the SEXIA project. This work has been supported by CICYT-SPAIN ALI-88-0187-CO2-02 and ALI-91-0786.

REFERENCES

1. Alonso, V., Aparicio, R. (1993).- "Characterization of European virgin olive oils using fatty acids".- Grasas y Aceites **44**, 18-24.
2. Aparicio, R. (1988).- "Characterization of foods by inexact rules: The SEXIA expert system".- J. Chemometrics **3**, 175-192.
3. Aparicio, R., Alonso, V. (1994).- "Characterization of virgin olive oils by SEXIA Expert System".- Prog. Lipid Res. **33**, 29-38.
4. Aparicio, R., Ferreiro, L., Alonso, V. (1994).- "Effect of climate on the chemical composition of virgin olive oil".- Anal. Chim. Acta 292, 235-241.
5. Aparicio, R., Ferreiro, L., Rodríguez, J.L. (1991).- "Caracterización de alimentos combinando reglas de decisión relacionales y lineales. Una aplicación al aceite de oliva virgen de Málaga".- Grasas y Aceites **42**, 132-142.
6. Aparicio, R., Ferreiro, L., Rodríguez, J.L. (1991).- "Characterization of Andalusian Virgin Olive Oils: SEXIA Project".- Andalusian Ministry of Agriculture, Seville, Spain.
7. Barranco, D., Rallo, L. (1984).- "The olive tree species harvested in Andalusia". - Instituto de Estudios Agrarios, Madrid, Spain.
8. Bisquerra, R. (1989).- "Introducción conceptual al análisis multivariante. Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD".- Promociones y Publicaciones Universitarias, Barcelona, Spain.
9. BMDP (1981).- "BMDP. Statistical Software".- University of California, Los Angeles.
10. Cadastro oleícola (1983).- Vol XVII- Laboratorios de Estudios Técnicos. (IROMA). Lisboa.
11. Derde, M.P., Buydens, L., Guns, C., Massart, D.L., Hopke, P.K. (1987).- "Comparison of rule-building expert systems with pattern recognition for the classification of analytical data".- Anal. Chem. **59**, 1868-1871.
12. Derde, M.P., Coomans, D., Massart, D.L. (1984).-, "SIMCA (Soft independent modeling of class anology). Demonstrated with characterization and classification of Italian olive oil".- J. Assoc. Off. Anal. Chem. **67**, 721-727.
13. EC. (1991). Diario Oficial. The Commission of the European Communities. Regulation nº2568/91, July 11th, 1991.
14. Ferreiro, L., Aparicio, R. (1992).- "Influencia de la altitud en la composición química de los aceites de oliva vírgenes de Andalucía. Ecuaciones matemáticas de clasificación".- Grasas y Aceites **43**, 149-156.
15. López Fernández, J., Tajadura Zapirain, J.A. (1992).- "AUTOCAD Avanzado. Versión 11".- Ed. McGraw Hill, Madrid, Spain.
16. M.A.P.A. (1988).- "El Olivar Español. Planes de reestructuración y reconversión".- Ministerio de Agricultura, Pesca y Alimentación. Dirección General de la Producción Agraria, Madrid, Spain.
17. Morettini, A. (1950).- "Olivicultura".- Ed. Ramo Editoriale degli Agricoltori, Roma, Italia.
18. Sabater, M.C., Boatella, J. De la Torre-Boronat, M.C. (1986).- "Application de l'analyse discriminante à la differentiation de huiles de differentes varietés".- Rev. Fr. Corps Gras **33**, 65-67.
19. Tsimidou, M., Karakostas, K.X. (1993).- "Geographical classification of Greek virgin olive oil by non-parametric multivariate evaluation of Fatty Acid composition".- J. Sci. Food Agric. **62**, 253-257.

(Recibido: Febrero 1994)