ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/239459050

Classification of olive oils from their fatty acid composition

Article · January 1983

CITATIONS
92

READS
1,449

3 authors, including:

Michele Forina
Università degli Studi di Genova
176 PUBLICATIONS **3,571** CITATIONS
SEE PROFILE

Sergio Lanteri
Università degli Studi di Torino
342 PUBLICATIONS **4,781** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Linking genetic resources, genomes and phenotypes of Solanaceous crops (G2P-SOL) View project

Project Developement of Machine Olfaction to Agricultural and Food Products View project

All content following this page was uploaded by Michele Forina on 17 February 2015.

The user has requested enhancement of the downloaded file.

CLASSIFICATION OF OLIVE OILS FROM THEIR FATTY ACID COMPOSITION

M.Forina, C.Armanino, S.Lanteri and E.Tiscornia
Institute of Pharmaceutical Sciences, viale Benedetto XV 3, I-16132 Genova (Italy).

SUMMARY

Some methods of graphic (eigenvector projection, non-linear mapping), parametric (linear discriminant analysis, Bayesian analysis) and non-parametric (KNN, learning machines, SIMCA) multivariate data analysis have been used for recognizing the geographical origin of olive oils from their fatty acid composition.

Three data sets have been tested: the first dataset, A, was of 572 oils from some regions of Italy (Liguria, Sardinia, Umbria, Apulia, Calabria, Sicily); the second, B, was of 188 samples from Italy, Portugal, Crete, Israel, Lebanon and Syria. Preliminary results from a third dataset, C, of 423 Portugese oils (production years 1976-78) are also reported.

All methods show high recognitive and predictive mean abilities (A: 92%, B: 86% correct class allocations) implying a high degree of dissimilarity between oils from different regions and great usefulness for characterization and control purposes.

Some regions were perfectly separated (e.g., A: West Liguria, Coast Sardinia, Inland Sardinia; B: Syria, Israel and Lebanon); in other cases a residual dissimilarity was found from misclassification matrices and Bayesian analysis. So, Sicilian oils show similarity with Calabria and South Apulia oils; Crete oils show decreasing similarity with oils from South Italy, East Liguria, Portugal and Syria.

The first principal component shows high correlation with latitude; in the case of Portugese oils, a strong correlation is obtained between latitude and a direction in the plane of the first two principal components. This direction does not change with production year and allows good separation of Douro river valley oils.

Loadings of variables on principal components or on separation hyperplanes, and SIMCA residuals enable us to evaluate

the variables of higher importance for class modelling and discrimination.

INTRODUCTION

Olive oil is still an important product for Mediterranean economics; moreover our tradition gives it a significance exceeding its nutritional value. Its fundamental economic importance in ancient times changed to religious significance joined with the cult of ancestors; consequently, the oil produced in a given country became a symbol of that country and of the family.

Many analytical data have been obtained on olive oils, particularly on the fatty acid composition of their lipid fraction. Sometimes original data are reported in the literature; otherwise the results of univariate statistical analysis are reported as histograms, means, ranges and standard deviations. In some cases statistical parameters show significant differences between oils from different regions; however, a useful part of the information in the chemical data has probably been lost, because the many intervariable correlations usually have not been taken into account.

The methods of multivariate data analysis allow us to extract the maximum information from analytical data. However, in spite of the widespread interest in food identity problems, only in a few cases have multivariate methods (linear discriminant analysis LDA, K-nearest neighbour KNN) been used in the data analysis of edible oils. Hartmann and Hawkes (1) used LDA to determine the geographical origins of peppermint oils from gas chromatographic data; Kacprzak and Higgins (2) used a method founded on the permitted range of fatty acid percentages and ratios for the automatic identification of edible oils.

Recently, the graphical methods of pattern recognition (3), some multivariate methods as LDA, KNN, learning machines, Bayesian analysis (4) and soft independent modelling of class analogy (SIMCA) (5,6) have been applied to Italian olive oils.

In this work we have continued our previous studies (4,5) and we have analyzed the data of fatty acid content of many samples of Mediterranean olive oils by some methods of pattern

recognition in order to:

- prove if more information useful to geographical origin identification can be extracted from analytical data;
- compare the data analysis methods as regards classification abilities, discriminatory power and computer time.

Some analytical data used here were obtained in our Institute (7), some others from the customs laboratory of Genoa (8,9) and from literature (10-18).

The measured differences between oils from various regions can be ascribed to many factors: olive variety, soil chemistry, treatments (manure, lopping, ploughing), diseases, climate and microclimate, olive harvesting and storage, olive pressing and oil storage, analytical procedure and operator errors. When we combine all these variation sources under the single name "geographical origin", we undoubtedly oversimplify, especially as regards procedure and operator errors. To minimize the consequences of these factors, the first stage of our investigation was the selection of homogeneous analytical data; so, a great part of the literature data was discarded. A further point was to establish whether the oil samples had been selected according to a sound statistical procedure, so that they are really representative of each geographical region; this is not true in some cases: e.g., the Israel and Lebanon oils used in this work had been sampled from stocks imported in Italy.

The analytical data collection we have used is therefore somewhat defective and the results of multivariate analysis must be regarded cautiously, however, the information gained can be valid because:

- the number of samples is high;
- in some cases data from different sources (Apulia) or from different years (Portugal) do not show source-dependent or year-dependent variations greater than region-attributed variations;
- in some cases (e.g., East and West Liguria, Crete) the analysis was made in the same laboratory, so that the dissimilarities can hardly be ascribed to the analytical procedure or to systematic operator errors.

Moreover, the results of multivariate analysis increase our

knowledge and agree with the information previously obtained by univariate statistical analysis.

DATA

Three data matrices have been used. The first one, *A*, was of I=572 rows (Italian oils) with K=8 columns (the variables were the percentages of k=1: palmitic, 2: palmitoleic, 3: stearic, 4: oleic, 5: linoleic, 6: linolenic, 7: arachidic and 8: eicosenoic acid). The second data matrix, *B*, was of I=188 rows (Mediterranean oils of which 106 where Italian oils randomly selected from the first matrix) with K=7 columns (column 8, eicosenoic acid percentages, was omitted since it was not reported in the available analysis of Portuguese oils). The thirs data matrix, *C*, was of I=423 rows (Portuguese oils) with K=7 acid percentage columns. The number of oil samples from each region in the three data matrices is shown in Table 1.

Table 1 - Structure of data matrices (number of oil datavectors in each class)

Matrix *A* Italian oils	(m)=class index	Region Class		Matrix *B* Mediterranean oils
25	(1)	North Apulia		
56	(2)	Calabria	South Italy (1)	33
206	(3)	South Apulia		
36	(4)	Sicily		
65	(5)	Inland Sardinia	Sardinia (2)	37
33	(6)	Coast Sardinia		
50	(7)	East Liguria		
50	(8)	West Liguria	North Italy (3)	36
51	(9)	Umbria		
Total 572				
		Lebanon 8		
		Israel 4	Israel and Lebanon (4)	12
			Crete (5)	20
			Syria (6)	20
Matrix *C* Portugal oils				
149	(1)	Portugal 1976/77		
121	(2)	Portugal 1977/78	Portugal (7)	30
153	(3)	Portugal 1978/79		
Total 423			Total	188

In the first analysis time Inland and Coast Sardinia oils of matrix *A* were considered as a single class; as were Ligurian and Apulian oils. They were separated after unsupervised graphical recognition and according to a well-known differentiation on the basis of sensory characteristics.

In data matrix *B*, Israel and Lebanon oils were put in a single class, because of the very low number of oil samples, after preliminary analysis had shown them to be very similar.

Italian and Portuguese oils were randomly selected from those used in matrices *A* and *C*, so that the number of samples was not much higher than that of Crete and Syria.

Data reported as "trace" in the original literature were processed as a random number in the range 0 - 0.1 %.

Data matrices are available from the authors upon request.

Each dataset was randomly subdivided into two datasets, training and evaluation sets; the percercentage of the datavectors in the training set was of between 65 and 75%. The subdivision was repeated ten times. The oils in the training set were used to generate classification rules; the performance of the rules was evaluated, first with the data in the training set, and measured as recognitive ability, then with the data of the evaluation set to test the predictive ability.

Let us indicate with *Y* the data matrix (*A* or *B* or *C*); Y_m indicates the datamatrix of region class m.

Data were usually scaled, so obtain an overall matrix *X* for which each fatty acid variable had mean zero and variance one and the M class matrices X_m :

$$x_{ik} = \frac{y_{ik} - \bar{y}_{.k}}{s_k} \qquad x_{ikm} = \frac{y_{ikm} - \bar{y}_{.k.}}{s_k}$$

where $\bar{y}_{.k} = \bar{y}_{.k.}$ is the mean value of variable k and s_k the estimated standard deviation.

Separate scaling has been used in SIMCA analysis (19):

$$x_{ik}^{(m)} = \frac{y_{ik} - \bar{y}_{.km}}{s_{km}}$$

where $\bar{y}_{.km}$ is the mean value of variable k and s_{km} its stan-

dard deviation estimate within the class m. The superscript (m) indicates that X results from one among the M possible separate scalings.

After the subdividing the objects in the training set into training and evaluation sets we obtained:

S overall variance-covariance matrix of X

S_m variance-covariance matrix of X_m

P pooled variance-covariance matrix

B matrix of the class means $b_{mk} = \bar{x}_{.km}$

C variance-covariance matrix of B

$R = P^{-1}C$

Sometimes data were multiplied by classification loadings obtained from the ratio of interclass to intraclass variance (for a two-class problem these loadings are equal to Fisher weights (20)).

METHODS

The following methods have been used:

EP : eigenvector projection

SNLM : simplified non linear mapping

LDA : linear discriminant analysis

MLDA : modified LDA

KNN : K-nearest neighbor

BA : Bayesian analysis

LLM : linear learning machines

SIMCA : soft independent modelling of class analogy

In EP (21) the sample points are projected from the k-dimensional space of the variables onto the plane of two principal components of matrix S or of one matrix S_m.

SNLM is a modification of non-linear mapping (22). In SNLM (3) the class barycenters b_m are represented by points in the representation plane in order to conserve as well as possible the interbarycenter distances in the k-dimensional variable space; then each object is mapped in order to preserve as well as possible the distances between the object and the three closest barycenters

LDA (23) computes Mahalanobis distances d_{im} of the object i from the M class barycenters as:

$$d_{im} = (x_i - b)\ P^{-1}\ (x_i - b_m)'$$

The object is classified into the class for which d_{im} is a minimum. From this equation, for each class pair mm', a decision surface can be obtained with the equation

$$f = w\ x' = 0$$

(where x' is the augmented vector, $x_{k+1} = 0$). The loading vector enables evaluation of LDA discriminatory importance of the variables for the separation of classes m and m'.

MLDA (24) project the objects on the principal eigenvectors of the matrix $R = P^{-1}\ C$ (the proiection on the two first eigenvectors is used also for display purposes). In this space the Euclidean distances of every object from the M class barycenters are computed, and the object is classified into the class of the closest barycenter. The number of eigenvectors to be used for the projection is the one that maximizes recognitive ability.

BA (23) computes the conditional probability density of the object i considered in the class m:

$$p_{im} = p\ (x_i/m) = (2\pi)^{-k/2} |S_m|^{-1}$$
$$\exp\left\{-1/2\ (x_i - b_m)\ S_m^{-1} (x_i - b_m)'\right\}$$

This conditional probability is multiplied by the a priori probability of class m, p_m, and by a loss factor, l_m, that is low when the damage caused by the erroneous assignment of an object to the class m is high

$$f_{im} = p_{im}\ p_m\ l_m$$

(in this work the a priori probabilities and loss factors have been assumed to be unity).

The object i is classified into the class with the highest value of the decision function f.

KNN (23) classifies the object in the evaluation set by means of a piecewise linear decision function. The distances of an object in the evaluation set from all the objects in the training set are calculated. The object in the evaluation set is classified in the class prevailing in the K nearest objects (here K is not the number of variables but an odd number, 3 or 5 generally; the symbol K was preserved here because it is always used in the relevant literature).

LLM (23) dichotomizes the hyperspace of variables so that a linear decision surface separates two groups of one or more classes. The loading w of the surface equation

$$f = w\, x' = 0$$

(where x' is the augmented vector, $x_{k+1} = 1$) is obtained by an iterative correction procedure with the datavectors of the training set.The correction procedure stops when all objects in the training set are correctly classified into one of the two class groups according to whether $w\, x'$ is positive or negative.

Repeated use of LLM gives a tree of binary decision makers. When the groups of classes are not linearly separable we used quadratic learning machines, where datavectors x are augmented by all variable-by-variable products, or a dampened linear learning machine, where the maximum allowed correction of the loading vector is gradually reduced.

SIMCA (19) is a rather peculiar method, as compared with those previously described, because of its ability to operate up to the upper levels of pattern recognition (25); outlier detection and class modelling are SIMCA facilities, before classification. Each class is first studied separately; a class model is built as a parallelepiped and its axes are the A_m significant eigenvectors of the matrix S_m (when $A_m = 1$ the model is a line segment, when $A_m = 2$ it is a parallelogram). The distance of an object from the model measures the degree of fit between the object and the model. Objects with a significantly high distance are discarded as outliers; a "SIMCA box" is built in the $(A_m + 1)$-dimensional space of the A_m eigenvectors plus the distance from the model, and this box has the class model as a base and the maximum permitted distance as height.

The loadings of the variables on the A_m eigenvectors allows us to evaluate the "modelling power" of each variable; the distance between two class models gives the interclass distance; the contribution of the variable k to this distance gives the discriminatory power of that variable.

Data analysis was performed with ARTHUR 81 (Infometrix Inc., Seattle,Wa.) and PARVUS (a package for desk computers that can be obtained from the authors). An Olivetti P6060 computer was used.

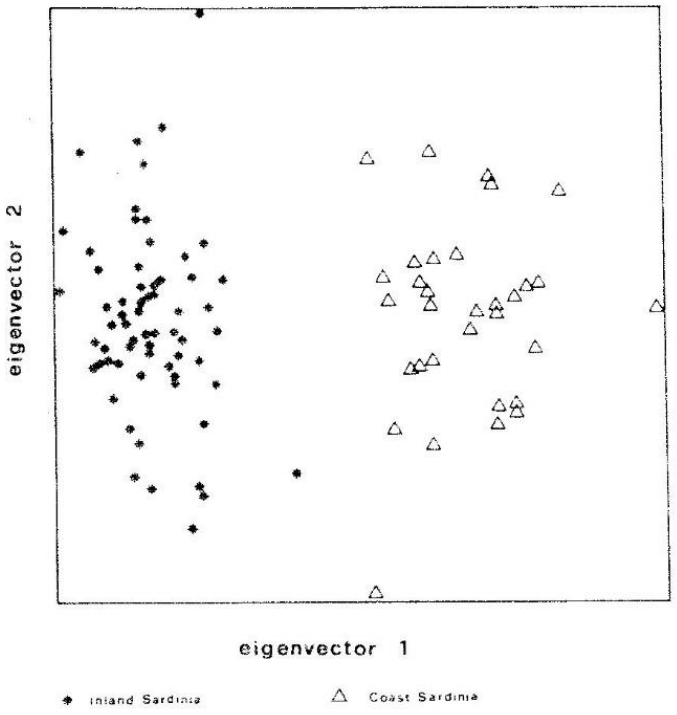

Figure 1 - Eigenvector plot of Sardinian oils

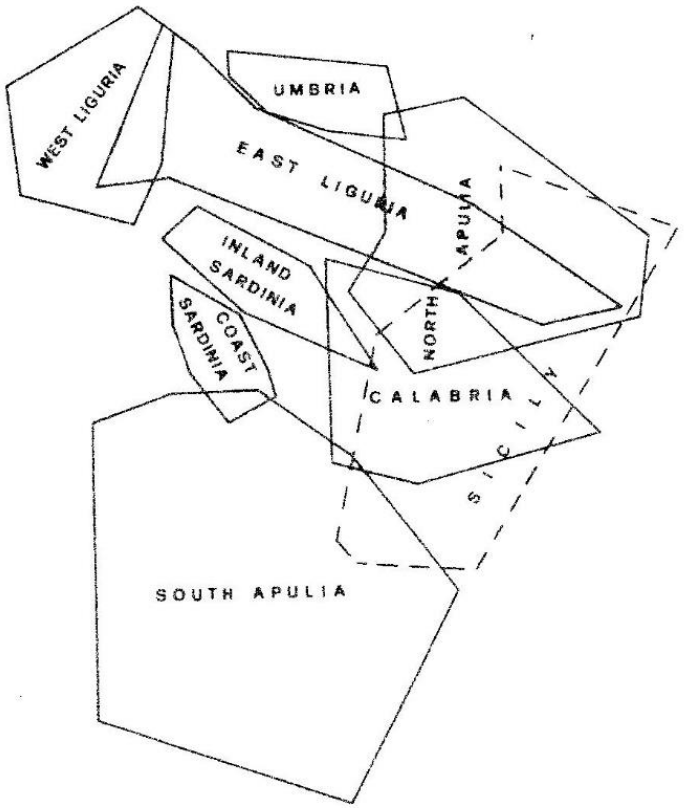

Figure 2 - Eigenvector plot of Italian oils (contour map)

RESULTS AND DISCUSSION

Graphical analysis

First we used EP on pairs of the first three eigenvectors in the analysis of the whole dataset or of some classes or of a class alone. Some EP's are shown in Fig.1-5. Because of the high number of samples, sometimes EP's are reported as contour maps. Graphical analysis shows that:

Matrix *A* (Figures 1 and 2)

- Sardinian oils separate into two distinct clusters and they are the Coastal and Inland Sardinian oils respectively; Apulia oils separate into North and South Apulian oil clusters.
- A good separation is obtained between East and West Liguria oils.
- No separations are detected within any other Italian region.
- In the plane of the first two eigenvectors, classes 2-3-5-6-7 8-9 seem well separated; class 1 partly overlaps 7, and class 4 overlaps 1, 2 and 7. The third eigenvector improves the separation of classes 1 and 7.

Matrix *B* (Figures 3,4 and 5)

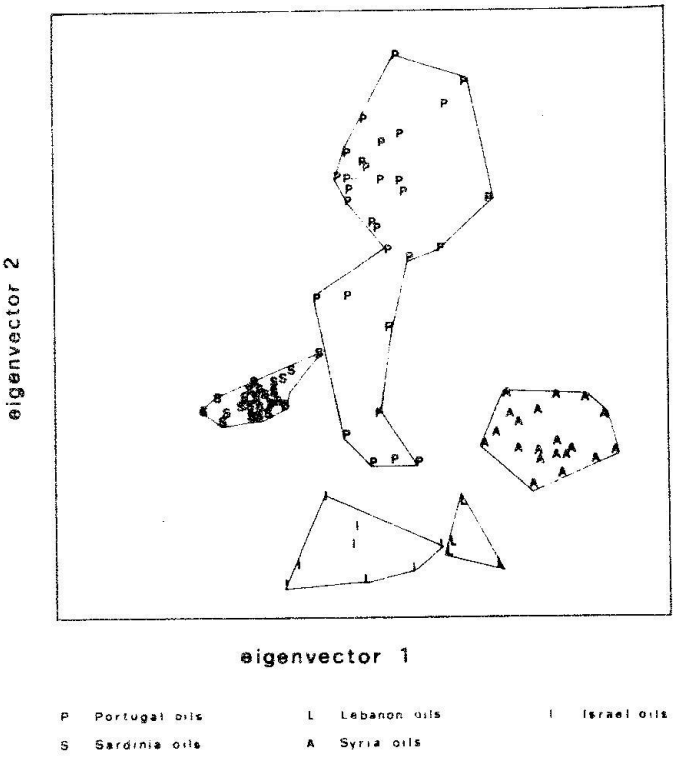

Figure 3 - Eigenvector plot of Mediterranean oils: Portugal, Lebanon, Syria, Israel and Sardinia

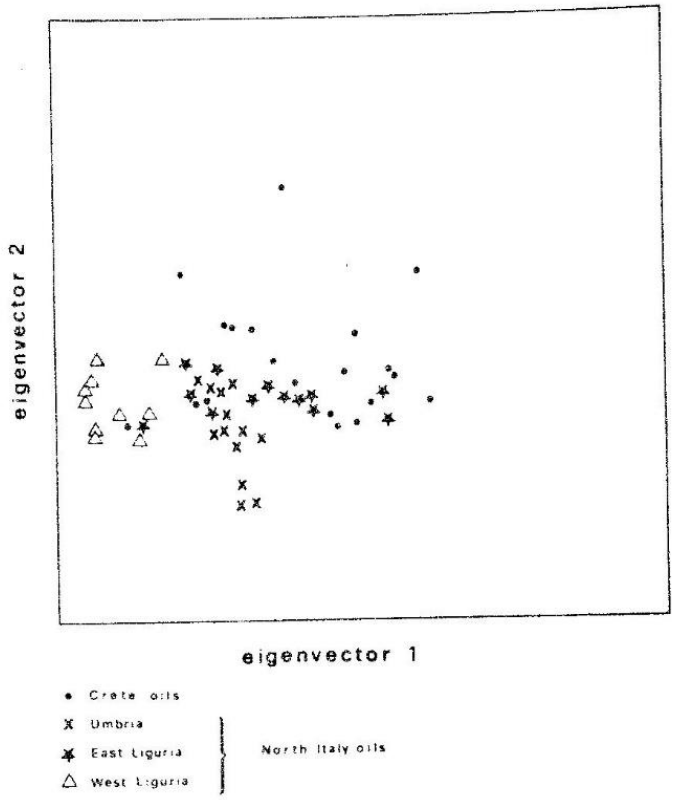

Figure 4 - Eigenvector plot of Mediterranean oils: Crete, North Italy oils

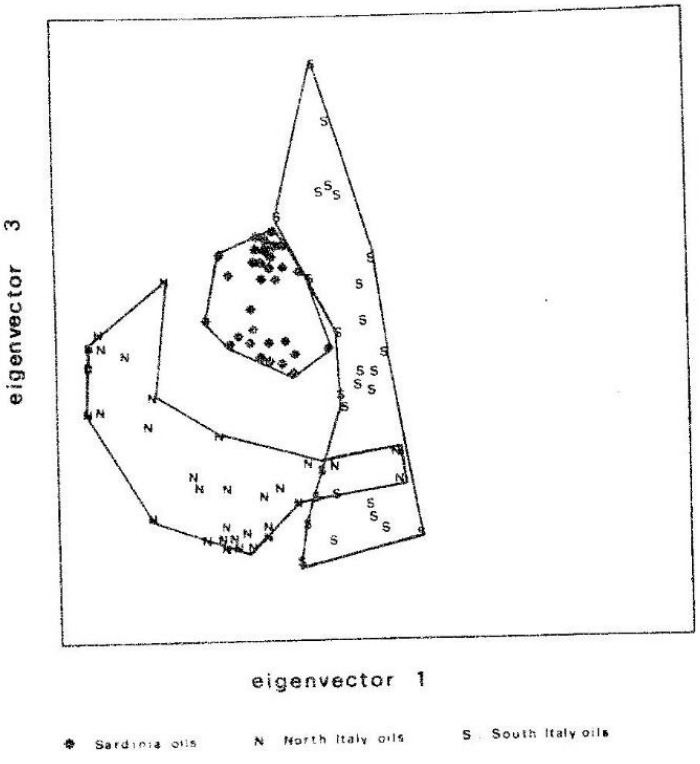

Figure 5 - Eigenvector plot of Mediterranean oils: Sardinia, North and South Italy oils

- No clusters are detected within any class.
- Israel and Lebanon oils are very close in the plot of the first two eigenvectors. We have therefore put these oils into a single class, because the number of samples is too low to justify the separate use of prediction methods.
- Sardinia, Portugal, Syria and Israel plus Lebanon are well separated.
- Crete oils overlap the Italian oil classes, and Portugal oils overlap South Italy oils (not shown in Figure 3).

Matrix *C* (Figure 6)

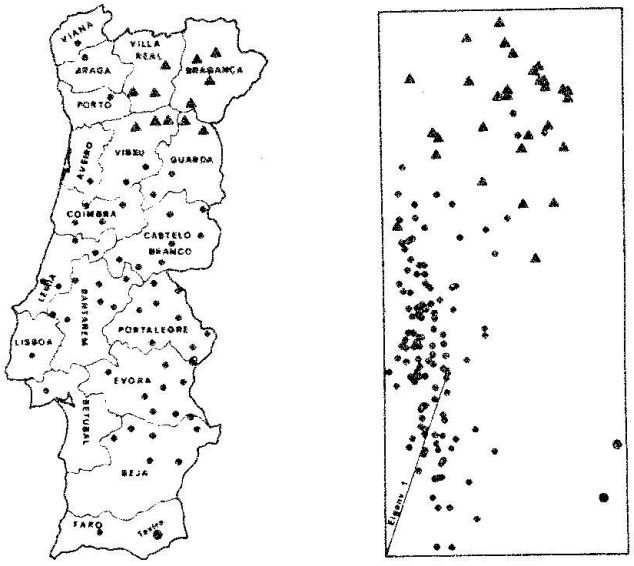

Figure 6 - Plot of Portuguese oils (1978/79) on the rotated plane of the first two eigenvectors: ▲Douro valley oils; ● Tavira; ✱ remaining Portuguese regions. At left the Portugal map with the sampling points.

In the plot of Portuguese oils we have rotated the axes in the plane of the two first eigenvectors so that the ordinate axis has the maximum correlation with the direction South-North.

- We see that two oils separate in the lower right corner; they are oils of Tavira, with very low oleic acid content and high percentage of linoleic acid; these two separate points are found again in Portugal 77/78 and Portugal 76/77, so Tavira

must be regarded as an atypical Portuguese region.
- in the upper part of the plot there are, well separated, the oils of the upper Douro valley.
- An almost vertical line can be seen with very high density of Center and South Portugal oils: this line can be one axis of the model of Portuguese oils.

A noticeable constancy in the overall characteristics of Portuguese oils is evidenced by the fact that the loadings of the variables on the first two principal components change little with year of production.

Besides EP, we used some simplified non-linear maps. Figure 7 shows SNLM of Italian olive oils. The dispersion polygons have been obtained by lines joining the points farthest from the barycenter with no more than one point being disregarded. In the point-by-point map, however, within the dispersion polygons some areas with greater point density can be seen. SNLM shows that the classes are more separated than appears from the overall eigenvector plot of Figure 2.

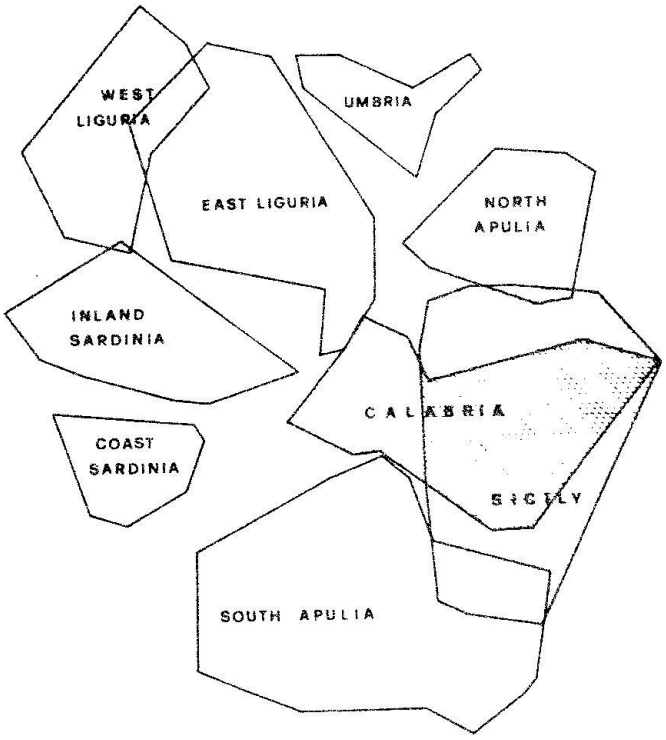

Figure 7 - SNLM of Italian oils

Table 2 - Results of classification analysis. Data matrix *A*. Results are reported as the percent recognitive (R) and predictive (P) abilities for the same random subdivision between training and prediction sets. KNN recognitive ability is 100 % because of method characteristics

Method:	KNN				LLM		LDA		MLDA		BA	
	(a)	(b)	(c)	(d)	R	P	R	P	R	P	R	P
Class												
1 North Apulia	100	88	83	75	100	100	100	88	100	92	100	38
2 Calabria	100	88	85	100	89	94	88	94	53	45	95	75
3 South Apulia	98	100	97	100	99	96	96	100	93	89	96	100
4 Sicily	60	53	38	60	77	20	76	80	77	29	100	87
5 Inland Sardinia	100	100	100	100	100	92	100	100	100	100	98	100
6 Coast Sardinia	90	100	100	92	100	100	100	100	90	100	100	100
7 East Liguria	84	100	81	94	94	99	91	100	94	81	100	100
8 West Liguria	100	100	86	100	100	100	100	100	97	100	100	100
9 Umbria	100	100	100	95	100	100	97	90	100	100	100	85
Total	95	94	89	94	97	91	95	96	90	83	98	92

KNN predictive abilities: (a) scaled data; (b) scaled and weighted; (c) non scaled; (d) eicosenoic acid omitted.

Classification methods

Tables 2 and 3 show the result of classification analysis; these results refer to one of the ten random subdivisions between training and evaluation sets, and they are very close to the mean of the ten subdivisions (except in the case of Bayesian analysis applied to North Apulia oils: mean predictive ability 79%).

Table 3 - Results of classification analysis
Data matrix *B*. (see Table 2 for symbols)

	Method: KNN		LLM		LDA		BA	
Class	(a)	(c)	R	P	R	P	R	P
1 South Italy	90	86	96	67	85	85	90	100
2 Sardinia	100	100	100	92	96	100	100	100
3 North Italy	80	100	91	86	84	82	96	73
4 Portugal	100	75	96	100	81	89	100	100
5 Israel and Lebanon	100	100	100	100	100	100	100	50
6 Crete	33	33	81	50	71	67	94	67
7 Syria	100	100	100	100	100	100	100	100
Total	89	86	95	86	88	88	97	90

Table 4 - KNN misprediction matrix. Dataset *A*. Data are percent attributions computed on the ten subdivisions between training and prediction set

Computed class	1	2	3	4	5	6	7	8	9
True class									
1 North Apulia	91	9							
2 Calabria		98	2						
3 South Apulia			97	3					
4 Sicily	16	20	12	52					
5 Inland Sardinia					100				
6 Coast Sardinia					5	95			
7 East Liguria							95	4	1
8 West Liguria								100	
9 Umbria									100

Table 5 - KNN misprediction matrix. Dataset *B*. (Percent attributions on ten random subdivisions)

Computed class	1	2	3	4	5	6	7
True class							
1 South Italy	84		6	1		9	
2 Sardinia		100					
3 North Italy	4		93	1		2	
4 Portugal	2	1		91		6	
5 Israel and Lebanon					97		3
6 Crete	20	19	31	2		28	
7 Syria							100

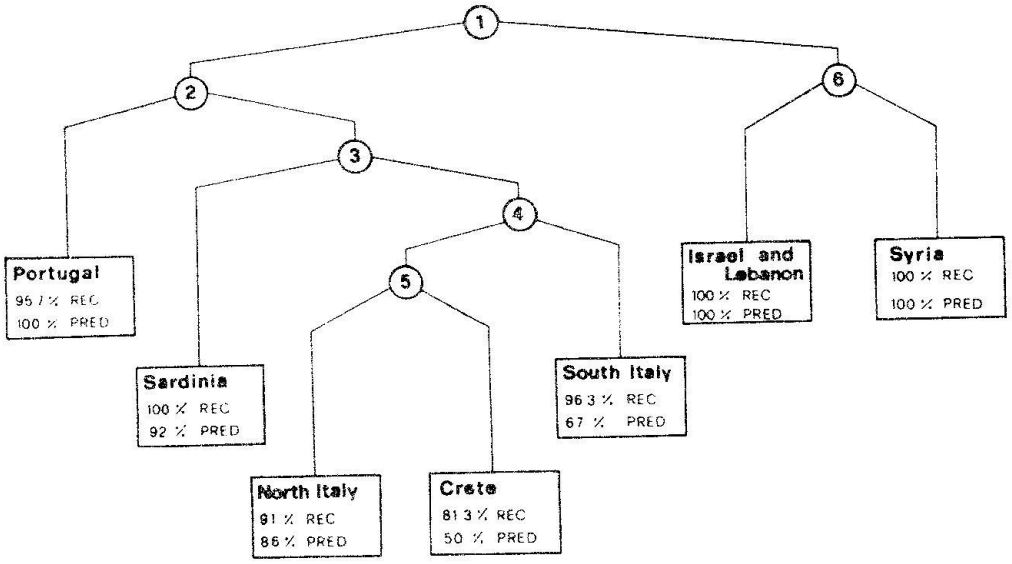

Figure 8 - Binary decision tree of linear learning machine. Dataset *B*.

Tables 4 and 5 show a more detailed analysis of KNN prediction results; Figure 8 shows the binary decision tree of learning machines applied to dataset *B*.

The high recognitive and predictive abilities shown by all the methods point out that the fatty acid composition is very useful to locate the geographical origin of an olive oil; in many cases of reduced attribution problems, when, for example,

the uncertainty is between two regions, the set of fatty acid percentages may be sufficient to give a reliable decision.

The increase of the useful (to classification) information extracted from the analytical data can be estimated by comparison, in the dataset *A*, between the multivariate Bayesian analysis (4% mean predictive error) and univariate and bivariate analysis. Oleic acid gives the best result in univariate analysis: 39% mean predictive error; the pair oleic-linoleic acid shows maximum predictive ability among the twenty-eight possible combinations: 14% mean predictive error.

Oleic and linolenic acids appear as the more important in the regional classification; owing to their high concentration, the analytical error is low and their discriminatory ability is not affected by the analytical method variance. Because of that, when KNN is applied to rough data, where the importance of variables grows with their numerical range, we have yet a high predictive ability; on the other hand, when we discard eicosenoic acid (its percentage is very low in all samples), predictive ability stays almost constant.

The use of scaled variables multiplied by classification loadings (Fisher weights) does not increase predictive ability; indeed, the classification loadings are of unambiguous effect in a two class problem. In a many class problem the loadings have a mean significance: their use can increase the distance, in the hyperspace of the new variables, of classes well separated in the space of the old unweighted variables at the expense of the distance of two ill-separated classes (e.g.,Sicily and Calabria).

The predictive abilities of each class and the repartition of prediction errors show a very good agreement with the graphical description of Figure 2 and of eigenvector plots. In the case of Italian oils, the wide superimposition of Sicily to Calabria and South Apulia explains the prediction errors; moreover, the few objects of Sicily are spread over a great range in the variable hyperspace, where objects of Calabria (prevailingly) are present too. Because of the higher number of Calabria objects,the probability that a Sicilian object has a Calabrian neighbour is higher than a Calabrian oil has a Si-

cilian one.

Instead, using Bayesian analysis, insensitive to the number of objects in the classes, when two class distributions are partly superimposed the class with more concentrated distribution will show a lower incorrect attribution percentage. With LDA the percentage of erroneous prediction will be equal for the two contiguous classes.

As Sicily in dataset *A*, in dataset *B* Crete shows a widespread distribution and a low object number; so KNN classifies many Cretan oils in other regions.

The use of geographical more-region classes of Italian oils in dataset *B* and the reduced object number increase prediction errors for Italian oils; in the eigenvector plots of Figures 3-5 a cross-road can be seen, where East Liguria, North Apulia, North Portugal and Crete oils meet or mix. This shows the inadequacy of a regional approach only, even though the administrative divisions of a country often reflect similar climatologic and geological characteristics and hystorical conditions favourable to the predominance of selected olive varieties and olive collection procedures.

As regards the performance of classification methods, LDA shows surprisingly high recognitive and predictive abilities, at least with dataset *A*, in spite of the very simple statistical model, where variance-covariance matrix is supposed class-independent, so that the class distributions appear as a series of equal confidence hyperellipsoids in the space of the variables, each class hyperellipsoid being centered in the corresponding class barycenter.

Really, we can see (Table 6) that many classes show high correlation coefficient of some pairs of variables: oleic-linoleic acid, palmitic-oleic acid; moreover, oleic and linoleic acid have great discriminatory importance (as shown by bivariate analysis). So, the pooled variance-covariance matrix gives a valuable approximation of single class distributions, at least in the directions of higher discriminatory importance.

The decision surface of LDA is a hyperplane located by the intersection of the confidence hyperellipsoids of two classes: the good performance of LDA explains that of linear lear-

ning machines.

Table 6 - Main correlation within variables. Data-set *A*. (The running indexes of the pair of variables with high correlation coefficient is reported with the sign of the correlation)

Class:	1	2	3	4	5	6	7	8	9
Pairs of variables with significant correlation	4,5-	1,4-	1,4-	1,4-		1,4-	4,5-		1,3+
	6,7+	4,5-	1,2-	2,4-		4,5-	4,7-		3,6+
	1,4-		4,5-	1,2+			6,7+		3,5-
	1,5+		2,4-	3,6+			4,6-		1,6+
				4,5-			5,6+		1,4-
				4,7-			5,7+		2,6-
				1,7+					
				3,5-					
				5,6-					

The highly negative correlation coefficient of oleic, linoleic and palmitic acid is not surprising: being oleic acid the first major oil component, the variations of its percentage must be reflected by the seconds (palmitic and linoleic) major components; the negative correlation between major components must be considered as a "standard characteristic" of olive oils. A class, therefore, can be characterized by the correlations between minor components or by a low negative correlation between major components; these low correlation may be due to low variance of oleic acid percentage, or to the fine class structure.

In the case of West Liguria, where the correlation coefficient between major components are low (although significant), Derde, Coomans and Massart (5) succeded, by SIMCA anlysis,in singling out two subgroups, the first where a diminution of oleic acid percentage strictly correlates with a higher percentage of palmitic acid, the second where oleic acid correlates with linoleic acid.

SIMCA analysis

SIMCA has been first used in data analysis of olive oils by Derde, Coomans and Massart (5). This method studies the similarities between the objects of a class to obtain a mathematical model of the class. This model can be seen as a box (SIMCA box) in the hyperspace of the variables. The final objective of SIMCA analysis on olive oils is to obtain well separated boxes for each region, so that an object within a box can be classified as "typical" of the corresponding region, and an object falling outside all class boxes must be regarded as an outlier, in our case an oil without an evident geographical origin or an adulterate oil.

Table 7 - SIMCA modelling power. Dataset A. For each class the running index of the three variables with the highest modelling power is reported

Class:	1	2	3	4	5	6	7	8	9
Variable									
First	4	4	4	4	4	4	4	2	1
Second	6	1	1	1	1	5	7	4	3
Third	3	6	2	2	6	3	3	5	6

Table 8 - SIMCA distances between classes. Dataset A

Class:	1	2	3	4	5	6	7	8	9	1st[&]	2nd[&]
1		2.5	4.5	2.2	6.0	8.2	3.8	5.8	6.9	4	2
2	2.5		1.5	1.1	4.4	5.4	3.2	5.4	17.7	4	3
3	4.5	1.5		1.6	4.9	4.4	4.3	4.0	36.9	2	4
4	2.2	1.1	1.6		6.0	7.0	5.0	6.8	21.2	2	3
5	6.0	4.4	4.9	6.0		2.4	4.6	4.4	28.9	6	2
6	8.2	5.4	4.4	7.0	2.4		6.8	5.4	41.3	5	3
7	3.8	3.2	4.3	5.0	4.6	6.8		3.2	7.2	8	2
8	5.8	5.4	4.0	6.8	4.4	5.4	3.2		16.5	7	5
9	6.9	17.7	36.9	21.2	28.9	41.3	7.2	16.5		1	7

& Running index k of the closest classes

Table 9 - Discriminatory importance of the variables. Dataset A. The running index of the three variables with the highest discriminatory importance as: a) Fisher weight; b) LDA loadings; c) SIMCA discriminatory power, is reported

Class	2			3			4			5			6			7			8			9		
	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
1	4	4	4	4	4	4	4	4	4	5	8	4	5	5	4	8	8	8	7	4	6	8	4	5
	1	5	1	2	2	2	1	5	2	4	5	5	4	8	5	1	1	4	8	8	7	3	5	4
	2	1	2	5	8	5	2	1	3	8	4	8	8	6	8	2	5	1	6	5	8	7	8	8
2				5	4	4	8	4	4	5	4	5	5	4	5	8	4	4	6	4	6	4	4	5
				2	5	5	7	5	1	8	5	1	6	5	1	4	5	8	7	5	7	1	8	4
				4	1	2	1	1	7	1	1	8	8	1	6	1	1	6	8	1	8	5	1	1
3							5	4	4	2	8	4	2	4	4	4	4	4	7	4	4	4	5	5
							8	5	7	8	4	1	8	5	1	5	8	5	6	5	6	5	4	4
							2	1	3	1	5	2	1	8	2	8	1	2	4	8	7	2	8	1
4										8	4	4	8	4	5	8	4	8	6	4	6	8	8	5
										6	5	8	6	5	4	6	8	4	7	5	8	7	5	4
										5	8	5	5	8	8	4	1	6	8	1	7	4	3	1
5													5	5	4	5	5	4	7	4	4	5	4	5
													4	4	5	4	1	5	6	7	5	4	5	4
													3	1	3	3	4	3	4	1	7	2	7	1
6																5	5	5	7	4	5	5	5	5
																4	1	4	4	7	4	4	4	4
																2	8	6	5	5	7	3	1	3
7																			7	5	7	3	4	5
																			5	4	6	4	5	4
																			6	7	5	2	1	3
8																						5	4	5
																						6	5	4
																						4	1	6

The SIMCA analysis here performed on oils of dataset A tries how much fatty acid percentages bring closer to the final objective. No analysis of data structure within the nine Italian classes was made. Results are reported in Tables 7-10 and in Figures 9-10. The final models of each region were obtained by discarding some outliers, which number is reported in Table 10.

Modelling and discriminatory powers of the variables confirm the high importance of major components. Discriminatory powers well agree with univariate classification scores and less with LDA loadings. The oversimplification of the "mixed" LDA model explains this disagreement: eicosenoic acid, e.g.,

shows high importance in the decision surface between classes 6 and 7; being eicosenoic acid at trace level for all oils in the two classes, LDA gives high classification importance to a random variable. Therefore, while univariate classification scores may be of great usefulness, mainly because of quick computing, LDA loadings must be used cautiously.

SIMCA interclass distances agree very well with the graphical representation of Figure 7. SIMCA classification ability was computed on the whole dataset, including the outliers discarded in the development of the model. Data in Table 10 were

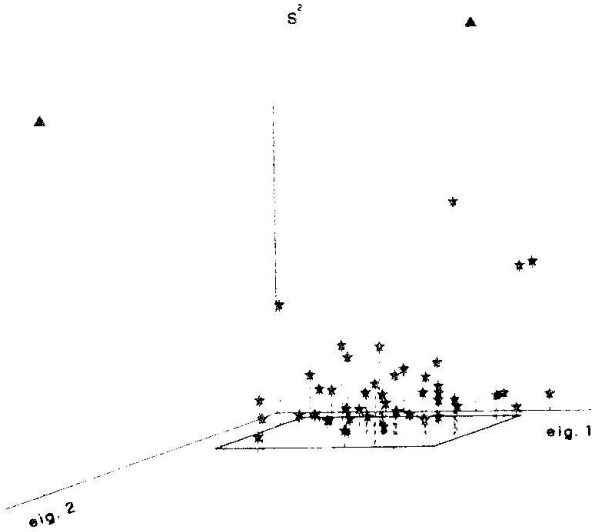

Figure 9 - SIMCA distances of East Liguria oils from their SIMCA model shown as a rectangle in the plane of the two components of the model

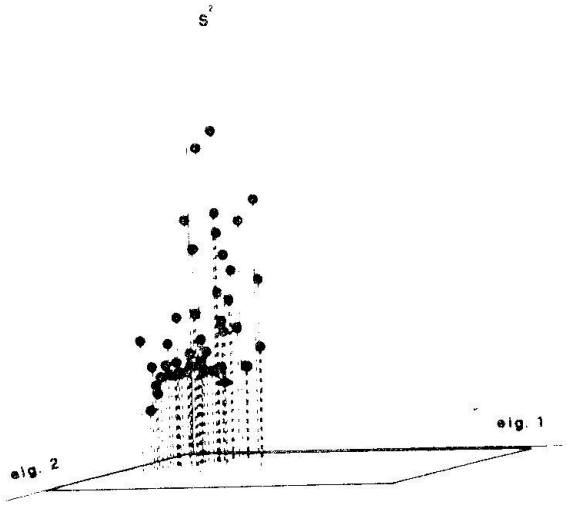

Figure 10 - SIMCA distances of Umbrian oils from the model of East Liguria oils

obtained by using SIMCA as a classification technique, i.e. a datavector was classified into the class with the lowest SIMCA distance.

Table 10 - SIMCA misclassification matrix. Dataset A

Computed class / True class	1	2	3	4	5	6	7	8	9
1 North Apulia	19		2	4					
2 Calabria		32	4	20					
3 South Apulia			195	11					
4 Sicily			1	35					
5 Inland Sardinia			4	3	58				
6 Coast Sardinia			2	1		30			
7 East Liguria							47	3	
8 West Liguria								50	
9 Umbria							7		44
Outliers	3	6	11	6	7	4	7	1	6
erroneous fit to the model	24	25	36	55	0	0	17	0	0

Classification ability is 89%, but rises to 95% when the outliers removed in the stepwise development of the class model are discarded (preliminary results, obtained with the variables weighted according to their analytical precisions or to the resulting residual standard deviations (19), show a better SIMCA classification ability).

In the last row of Table 10 we report the number of datavectors that erroneously fit the class models. A datavector may be counted here even if it is closer to another class. So we have 24 datavectors that erroneously fit the model of the first class, but classification analysis shows that they all fit better another class model, so that no datavector is erroneously classified into class 1.

The number of objects that erroneously fit to class models is very high in the case of South Italy oils, both for the low interclass distances and the high object number of South Apulia. Inland and Coast Sardinia, West Liguria and Umbria are impene-

centages could be used in multivariate analysis; we believe that accurate models of classes can be obtained, provided that botanic, climatic and geochemical informations are used to select suitable oil regions.

REFERENCES

1) Hartmann,N. and Hawkes,S.J. (1970) Statistical analysis of multivariate chromatographic data on natural mixtures, with particular reference to peppermint oils. J.Chromatogr.Sci.,8,610-611

2) Kacprzak,J.L. and Higgins,V.R. (1979) A computer program for the automatic identification of edible oils by gas-liquid chromatography. Anal.Chim.Acta,112,443-448.

3) Forina M. and Armanino C. (1982) Eigenvector projection and simplified non linear mapping of fatty acid content of Italian olive oils. Annali di Chimica,72,127-141.

4) Forina M. and Tiscornia,E. (1982) Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content. Annali di Chimica,72,143-155.

5) Derde,M.P.,Coomans,D. and Massart,D.L. (1982) Characterization and classification of Italian olive oil with SIMCA. submitted to J.Agric.Food Chem.

6) Derde,M.P.,Coomans,D. and Massart,D.L. (1982) Effect of scaling on class modelling with SIMCA. submitted to Anal. Chim.Acta.

7) Tiscornia,E. Unpublished data.

8) Paganuzzi,V. (1975) The composition of Israel olive oil. (Italian) Riv.Ital.Sostanze Grasse,52,302-306.

9) Paganuzzi,V. and Leoni,E. (1975) The composition of Lebanon olive oil.(Italian) Riv.Ital.Sci.Alimentazione,4,269-271.

10) Cotichelli,O. and Petruccioli,G. (1968) The physico-chemical characteristics of Syria virgin olive oils.(Italian) Industrie Agrarie,6,408-415.

11) Portugal Institute for Oil and Oil Products (1978) X Oil Cadastre (Portuguese).

12) Portugal Institute for Oil and Oil Products (1980) XII Oil Cadastre (Portuguese).

13) Alamanni,U.,Bonfigli,A. and Saba,A. (1968) Chemical and physico-chemical characteristics of Sardinian olive oils. (Italian) Boll.Lab.Chim.Prov.XIX,255-271.

14) Doro,B. and Remoli,S. (1969) Physico-chemical characteristics and acidic composition of some virgin South-Italy olive oils. (Italian) Riv.Ital.Sostanze Grasse,46,467-477.

15) Cucurachi,A. (1967) Physico-chemical characteristics and fatty acids of Calabrian olive oils.(Italian) Riv.Ital. Sostanze Grasse,44,172-177.

16) Cucurachi,A. (1964) Fatty acid composition of Apulian olive oils.Note I. Variety "Coratina".(Italian) Riv.Ital.Sostanze Grasse,41,234-242.

17) Cucurachi,A. (1967) Fatty acid composition of Apulian olive oils. Note II. Variety "Cima di Mola".(Italian) Riv. Ital.Sostanze Grasse,44,260-266.

18) Petruccioli,G. (1966) Physico-chemical characteristics of Umbrian virgin olive oils. (Italian) Olearia,20,24-29.

19) Wold,S. et al. (1981) Pattern recognition of disjoint principal components models (SIMCA).Philosophy and methods. Symposium on Applied Statistics, Copenhagen, January 22.

20) Kowalski,B.R. (1977) Chemometrics: theory and application. ACS Symposium Series 52, Am.Chem.Soc.,Washington D.C.,43.

21) Kowalski,B.R. (1975) Measurement analysis. Anal.Chem.,47, 1152A-1162A.

22) Kowalski,B.R. and Bender,C.F. (1973) Pattern recognition. II. Linear and nonlinear methods for displaying chemical data. J.Am.Chem.Soc.,95,686-692.

23) Tou,J.T. and Gonzales,R.C. (1974) Pattern recognition principles. Addison-Wesley Publ.Co.Inc., Massachusetts.

24) Kawahara,F.K. and Yang,Y.Y. (1976) Systems chemical analysis of petroleum pollutants. Anal.Chem.,48,651-655.

25) Albano,C. et al. (1978) Four levels of pattern recognition. Anal.Chim.Acta,103,429-443.

View publication stats