Pattern Recognition Analysis Applied to Classification of Wines from Galicia (Northwestern Spain) with Certified Brand of Origin

María J. Latorre,† Carmen García-Jares,† Bernard Mèdina,‡ and Carlos Herrero*,†

Departamento de Química Analítica, Nutrición y Bromatología, Facultad de Ciencias de Lugo,
Augas Férreas s/n, 27002 Lugo, Spain, and Laboratoire Interrègional de la Concurrence,
Consommation et de la Répression des Fraudes, 33405 Talence Cédex, France

In 42 white wines from Galicia (northwestern Spain) some trace elements were determined by atomic spectroscopy. Data were processed using multivariate chemometric techniques involving cluster analysis, principal component analysis, discriminant analysis, *K* nearest neighbors, and soft independent modeling of class analogy to develop a typification for wine samples of Rías Baixas origin. The wines with Certified Brand of Origin Rías Baixas can be differentiated from wines of the other two Certified Brands of Origin from Galicia, Ribeiro and Valdeorras, which are possible substrates for falsification. Using lithium and rubidium as key features, a nearly correct classification was achieved. The probability of a non-Rías Baixas wine being accepted as genuine is practically nil.

Keywords: *Wine, metals, trace elements, pattern recognition*

INTRODUCTION

Classification of product brands and quality of origin is a very active area for the application of chemometric classification procedures (Forina and Lanteri, 1984). The use of specific sensors for characterizing foodstuffs is being replaced by a trend to draw on the wealth of information available from the data provided by current analytical instrumentation. The extraction of useful information from an amount of data and the optimum use of this analytical information are important objectives of chemometrics.

The application of pattern recognition techniques for the classification of wines and other alcoholic beverages has widely increased in recent years. This implies objective methods based upon differences in composition. A number of examples were reported in the literature. Classification of wines of *Vitis vinifera* cv. Pinot Noir from France and the United States was achieved according to their geographic origins on the basis of gas chromatographic data of volatile compounds and mineral and trace element data (Kwan and Kowalski, 1978, 1980; Kwan et al., 1979). Different wine brands were distinguished using linear discriminant analysis and quadratic discriminant analysis (Armanino et al., 1990). Pattern recognition techniques were applied to data obtained from free amino acid profiles of 42 Portuguese wines; the characterization of wines to the corresponding original grape varieties in all cases studied was achieved (Vasconcelos and Chaves, 1989). Multivariate analysis of mineral and trace element data was effective in determining the geographic origin of different Spanish wines (Herrero and Mèdina, 1990; Latorre et al., 1992). Differentiation among German white wines from two different regions was performed by using linear discriminant analysis and *K* nearest neighbors (KNN) applied to volatile compound, nonvolatile acid, and amino acid data (Maarse et al., 1987). Typification of alcoholic distillates from Galicia by applying classification (cluster analysis and KNN) and modeling techniques (Bayes and partial least-squares) to data from chromatographic analyses was reported (Cruz Ortiz et al., 1993).

The purpose of this study is to differentiate between wines from brand (Denominación de Origen; Certified Brand of Origin) Rías Baixas and wines from the other two brands from Galicia, Ribeiro and Valdeorras, which can be used as possible substrates for falsification due to their similar organoleptic properties and color. Moreover, wines of Ribeiro and Valdeorras origin are sold at lower prices in the market than Rías Baixas wines. The basis for this differentiation is the content of some selected trace elements as determined by atomic spectroscopy.

EXPERIMENTAL PROCEDURES

Wine Samples. The most important criterion in the study of wine authenticity is that there should not be any doubt as to the geographic origin and the grape species from which they have been made. To be quite certain in this respect, the 42 wine samples representing the three production areas for this work were collected as follows: 21 wine samples from Rías Baixas were provided by the Certification Origin Council in the production area; 11 wine samples from Ribeiro and 10 from Valderorras were collected by the authors in the original areas directly from the producers. All of the wines came from unsuspected origin and were made with the traditional varieties for these Spanish wine-producing regions during the period 1990–1992. Samples were collected in 750-mL glass bottles and stored at 3–4 °C before analysis.

Analytical Procedures. According to previous works (Herrero and Mèdina, 1990), seven selected metals, Li, Rb, Na, K, Mn, Fe, and Ca, were measured in all wines using a Perkin-Elmer 2280 atomic absorption spectrometer. Li, Rb, Na, and K were determined by atomic emission spectroscopy and Mn, Fe, and Ca by atomic absorption spectroscopy according to the methods of the Office International de la Vigne et du Vin (OIV, 1978). Experimental conditions and sample dilution or additions are given in Table 1. All determinations were made twice.

Data Analysis. Each wine sample (object) was considered as an assembly of seven variables represented by the chemical data. These variables, called "features", formed a "data vector" which represented a wine sample. Data vectors belonging to the same group, such as geographic origin, were analyzed. The group was then termed a "category". Pattern recognition tools used in this work were as follows.

Autoscale. This is the most widely used scaling technique (Kowalski and Bender, 1972). The procedure standardizes a

† Facultad de Ciencias de Lugo.

‡ Laboratoire Interrègional de la Concurrence.

0021-8561/94/1442-1451$04.50/0

Table 1. Conditions Applied in AES and AAS Determinations

element	mode	wavelength (nm)	sample dilution
Li	AES	670.8	without dilution
Na	AES	589.4	1:10
K	AES	766.4	1:20
Rb	AES	780.0	without dilution
Ca	AAS	422.7	sample diluted 1:25 with $LaCl_3$ solution
Fe	AAS	248.3	without dilution
Mn	AAS	279.5	without dilution

variable k according to

$$y_{ik} = (x_{ik} - \bar{x}_k)/s_k$$

where y_{ik} is the value i for the variable k after scaling, x_{ik} is the value i for the variable k before scaling, $\bar{x}_k$ is the mean of the variable, and s_k is the standard deviation of the variable. The result is a variable with zero mean and a unit standard deviation.

Fisher Weights. This is a quantitative estimate of the utility of a given measurement for separating categories (Meloun et al., 1992a). The Fisher weight is the ratio between the square of the difference between the category means and the sum of the squared category standard deviations.

Cluster Analysis. Clustering techniques comprise an unsupervised classification procedure that involves a measurement of either the distance or the similarity between objects to be clustered. Objects are grouped in clusters in terms of their nearness or similarity. The initial assumption is that the nearness of objects in the p-space defined by the variables reflects the similarity of their properties (Massart and Kaufman, 1983).

Principal Component Analysis (PCA). This procedure (Mardia et al., 1979) was used mainly to achieve a reduction of dimensionality, i.e., to fit a j-dimensional subspace to the original p-variate ($p > j$) space of objects and permit a primary evaluation of the between-category similarity.

Linear Discriminant Analysis (LDA). This classification procedure (Wold et al., 1984) maximizes the variance between categories and minimizes the variance within categories. The method renders a number of orthogonal linear discriminant functions, equal to number of categories minus 1.

K Nearest Neighbor (KNN). This classification method, which utilizes the distance between objects in the p-space as its criterion (Wold et al., 1984), is used to classify an object in the category which contributes the greatest number of K nearest known objects. It is a nonparametric method inasmuch as it does not formulate a hypothesis on the distribution of the variables used. Only the closest K objects are used in making any given classification. The importance of a given feature in making the decisions is proportional to its contribution to the distance calculation. The inverse square of Euclidean distance was used in this work.

Soft Independent Modeling of Class Analogy (SIMCA). This classification procedure uses linear discriminant functions derived from disjointed principal component analysis of the data (Wold, 1976). One set of functions is derived for each category studied by computing the category mean and a specified number of the principal components. Objects are classified into the category whose principal component model best reproduces the data. Only data points that are members of a given category are used in determining the model functions for that category. The importance of each feature in classification is determined by its contribution to the category covariance matrices.

The data analysis was performed in few steps:

(1) Preliminary data analysis by cluster and principal component analysis used the complete data set.

(2) Classification techniques LDA, KNN, and SIMCA were applied to the complete data set with a category arrangement: category 1, training set of 11 Ribeiro wines and 10 Valdeorras wines (21 non-Rías Baixas wines); category 2, training set of 21 Rías Baixas wines.

(3) For practical reasons it is important to know the minimum number of features needed to obtain a correct classification. This could be achieved by choosing features that contained the most discriminant information for the classification. The criterion used for selection was Fisher weights. Li and Rb were selected as key features.

Table 2. Trace Elements in Wines from Galicia[a]

element	mean	SD	max	min
	Non-Rías Baixas Wines (21 Samples)			
Li	32.0	11.6	58.0	12.0
Na	29.0	16.4	57.5	11.0
K	978	346	1430	340
Rb	1.0	0.7	2.5	0.1
Ca	89.6	16.9	118.5	64.2
Fe	7.9	3.8	17.4	1.5
Mn	2.3	1.1	4.5	1.0
	Rías Baixas Wines (21 Samples)			
Li	7.6	6.3	32.0	2.0
Na	28.9	25.7	103.5	10.3
K	620	161	898	407
Rb	2.4	0.6	3.6	0.7
Ca	83.9	19.9	128.3	49.2
Fe	3.3	3.2	13.8	0.5
Mn	1.3	0.5	2.6	0.7

[a] All results are in milligrams per liter except Li, which is in micrograms per liter.

(4) The reliability of the classification obtained before was checked. The 42 objects were randomly divided between training (or learning) set and evaluation (or prediction) set. KNN and SIMCA were applied on the basis of only two features selected in step 3.

Pattern recognition analyses were performed by means of the statistical software packages Statgraphics (Statgraphics, 1991) and Parvus (Forina et al., 1988) on a Gulf-Tech 486/33 computer using a Hewlett-Packard Laserjet II as graphic output.

RESULTS AND DISCUSSION

The results of the instrumental analysis of seven selected metals for Ribeiro and Valdeorras wines have been published elsewhere (Herrero and Mèdina, 1990), and the levels obtained were similar to those found by other authors in wines from Galicia (Fernández et al., 1987). In this work, we present results of additional determinations of some metals in Rías Baixas wines together with data analysis. A summary of the data is given in Table 2.

The search for natural groupings among the samples is one preliminary way to study the data structure. Cluster analysis describes the nearness between wine samples (objects). In this case, a matrix consisting of the squared Euclidean distances between objects was used as a similarity matrix. Thus, a similarity matrix $S_{42\times42}$ was constructed from the autoscaled data; the elements of this matrix were the squared Euclidean distances of one object from the rest. To obtain clusters, the Ward method was used. This agglomerative method considers in each step the heteregeneity or deviance (sum of squares of the distance of an object from the barycenter of the cluster) of every possible cluster that can be created by linking two existing clusters (Meloun et al., 1992b). The results obtained showed the presence of wine clusters; the data of mineral and trace element composition of wines contained useful information to achieve a two-category classification between Rías Baixas brand and non-Rías Baixas brand. The results of the cluster analysis are shown as a dendogram in Figure 1. At a similarity level of 0.7, six clusters were found, which can be identified as follows. The first cluster consisted of a series of six wines of Ribeiro origin plus one Rías Baixas wine. The second cluster was composed of seven Valdeorras wines. The third cluster was made up of four wines from Ribeiro plus two wines of Valdeorras origin. The fourth cluster was composed of two samples from Rías Baixas. The fifth cluster contained two wines from Rías Baixas and one wine from Valdeorras.

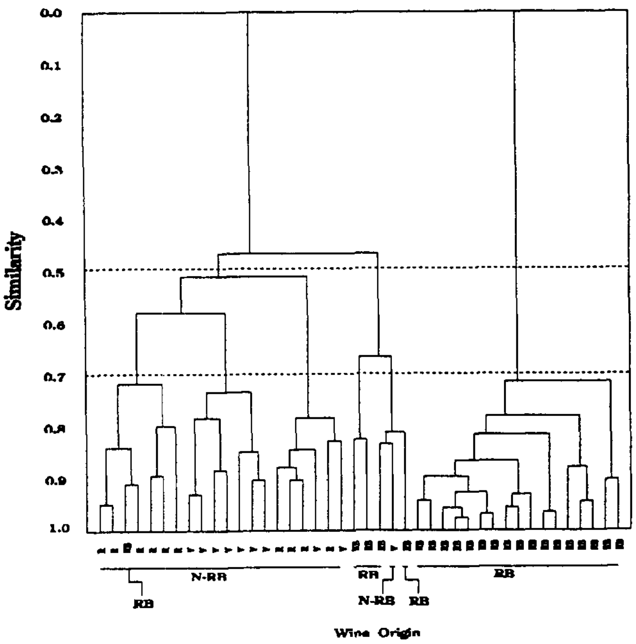

Figure 1. Dendogram of cluster analysis of wines of Rías Baixas origin (RB) and non-Rías Baixas origin (NRB). Wine codes: RB, Rías Baixas; R, Ribeiro; V, Valdeorras.

Table 3. Loadings of the First Two Principal Components

	variable						
	Li	Rb	K	Fe	Na	Ca	Mn
principal component 1	0.51	−0.39	0.39	0.48	0.17	0.14	0.37
principal component 2	0.27	0.52	−0.39	−0.12	−0.44	−0.40	−0.35

Finally, the sixth cluster included 17 wines of Rías Baixas origin. Examination of the dendogram at a similarity level of 0.5 showed three clusters: the first was formed by 19 wines of non-Rías Baixas origin and 1 Rías Baixas wine; the second cluster contained 4 wines from Rías Baixas plus 1 non-Rías Baixas; the last of the three clusters was composed of 17 wines of Rías Baixas origin. In this case, five wines of Rías Baixas origin were grouped in a cluster with a high level of similarity with non-Rías Baixas wines.

Principal components were calculated by using a routine of Statgraphics. Principal components are orthogonal, and each principal component is a linear combination of the original variables. From the coefficients (loadings) of features in the first and second principal components (see Table 3), lithium is the dominating feature in the first principal component (39.8% of the total variability), while rubidium dominates the second principal component (17.4% of total variability). The first three principal components, which account for 72.0% of the total variability, were considered to be sufficient for such data. Group classification by PCA afforded interesting results. When a three-dimensional plot of the objects in the space defined by the three principal components was drawn, a natural separation of the objects (wine samples) into two groups was achieved (Figure 2). Wines of Rías Baixas origin form a separate and homogeneous group. Wines of non-Rías Baixas origin appear in a less homogeneous group since this group is formed by wines from Ribeiro and Valdeorras brands. In this factor space, non-Rías Baixas makes up a group that includes three Rías Baixas wines. This result is consistent with the conclusions obtained by cluster analysis, where five wines of Rías Baixas origin were clustered with wines from non-Rías Baixas brands.

Three classification methods, LDA, KNN, and SIMCA, were applied to the complete data set (step 2) after autoscale to eliminate the effect of different size variables. Linear discriminant analysis was applied to an initial matrix containing the 42 objects and the 7 variables divided between Rías Baixas and non-Rías Baixas origin. The recognition ability for the two classes was highly satisfactory; all samples were correctly classified (see Table 5). Results obtained by KNN in the complete data set using inverse square of the Euclidean distance and $K = 4$ was less successful, with a correct classification of 93.8% of wines. Three wines from Rías Baixas were misclassified. The same results were obtained when KNN with $K = 5$ was applied. Using SIMCA, 95.2% of the wine samples were assigned to their production area (2 samples of Rías Baixas origin were misclassified). In no case were any of the non-Rías Baixas samples classified as Rías Baixas.

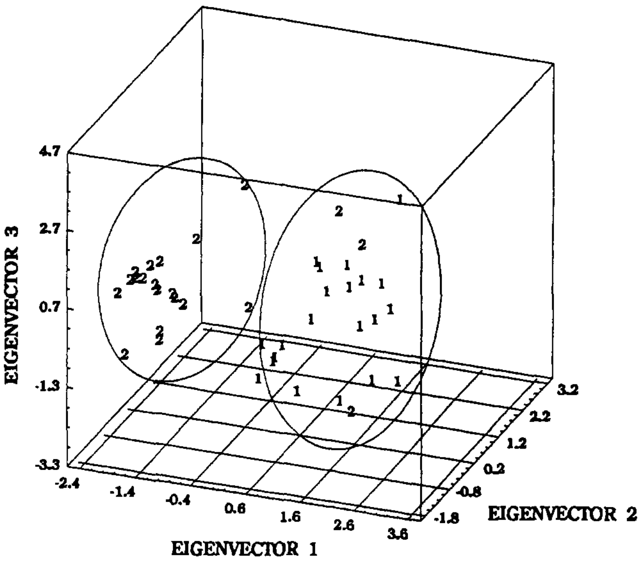

Figure 2. Eigenvector projection of wine samples. 1, non-Rías Baixas origin; 2, Rías Baixas origin.

Table 4. Fisher Weights of the Seven Elements Determined

element	Fisher wt	element	Fisher wt
Li	3.42	Mn	0.68
Rb	2.31	Ca	0.05
K	0.88	Na	0.00
Fe	0.86		

In step 3, a selection of minimum number of features for a correct classification was performed. Selection of a small number of key features offered other advantages besides increasing the reliability of mathematical classification results. Two-dimensional plots of the key features, one vs the other, allowed visual examination of the data set. The criterion for variable selection was the Fisher weight, and the results obtained are summarized in Table 4. Li and Rb were found to be the features that contained the most discriminatory information for classification. These results agree with the conclusions obtained by PCA, where Li and Rb were the dominant features in principal components 1 and 2. Features chosen after these two had much lower Fisher weights, showing that they contained insignificant amounts of additional information for the typification of wines from Galicia. Feature selection provided results analogous to those published by Maarse (Maarse et al., 1987), who found that German wines from the Rhein-Pfalz and Mosel regions could be distinguished by Rb, Na, Fe, and Li as key features. Also, Li and Rb were effective features for the classification of French wines of Médoc and Saint-Emilion origins (Lacasta, 1982; Etievant et al., 1988).

Table 5. Classification with LDA, KNN, and SIMCA

	LDA (All Features)		
category	non-Rías Baixas	Rías Baixas	% correct classification
non-Rías Baixas	21	0	100
Rías Baixas	0	21	100

	SIMCA (Two Selected Features: Li and Rb)	
category	recognition ability (%)	prediction ability (%)
non-Rías Baixas	99.4	99.0
Rías Baixas	94.4	90.0

	KNN (Two Selected Features: Li and Rb)					
	$K = 2$		$K = 3$		$K = 4$	
category	recognition ability (%)	prediction ability (%)	recognition ability (%)	prediction ability (%)	recognition ability (%)	prediction ability (%)
non-Rías Baixas	98.1	99.0	98.1	99.0	99.4	99.1
Rías Baixas	96.2	92.0	96.2	92.0	96.2	92.1

Finally, the reliability of the classification was tested (step 4). The 42 objects were randomly divided between training (or learning) set and evaluation (or prediction) set. The percentage of objects placed in the evaluation set was 25%. Such division allows us to have a sufficient training set as well as an evaluation set containing a representative number of samples. To obtain a good evaluation of recognition and prediction ability of each method using the two selected features, the previous division procedure was repeated 10 times for different constitutions of the two sets. In this case (Table 5), all classification methods give similar results. A high level of correct assignation of wines from non-Rías Baixas with a percentage of successes in recognition and prediction above 99% was achieved using KNN and SIMCA. For wines of Rías Baixas origin the percentage of correct classification was less successful. KNN and SIMCA provided a correct recognition above 95% and a correct prediction above 91%. This fact indicates that the pattern recognition procedures are selective for the non-Rías Baixas wines; the probability of a non-Rías Baixas wine being categorized as genuine Rías Baixas is virtually nil. However, a minor level of hits in classification and prediction of Rías Baixas origin suggests that there is a certain probability that a genuine Rías Baixas wine might be classified as non-Rías Baixas. These results aggree with the ones obtained by PCA and cluster analysis, where three and five samples from Rías Baixas were grouped, respectively, as being of non-Rías Baixas origin. The data were then examined by a two-dimensional plot of the two key features. A plot of lithium vs rubidium is shown in Figure 3; one sample of genuine Rías Baixas brand was classified as false, confirming the possibility of rejection as noted earlier.

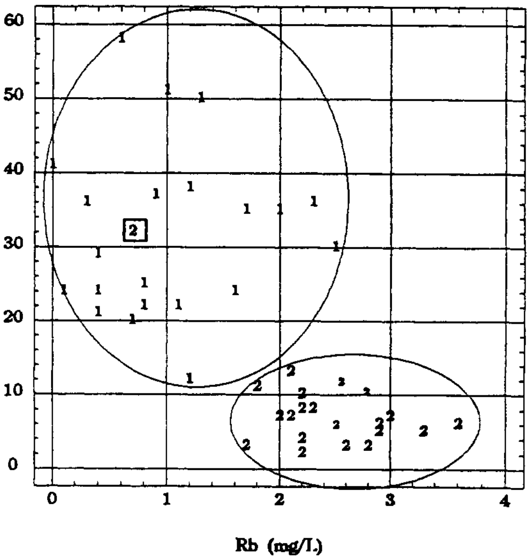

Figure 3. Plot of first selected feature, lithium, vs second selected feature, rubidium. 1, non-Rías Baixas origin; 2, Rías Baixas origin.

CONCLUSION

The content of selected mineral and trace elements of wines from Galicia was used to differentiate between wine samples of Rías Baixas origin and other wine samples from Galicia corresponding to other brands. PCA and cluster analysis revealed the occurrence of groupings between the analyzed samples according to their brand. The content of mineral and trace elements in wine may be influenced by a few factors, such as the level of these elements in soil, fertilizing practices, and processing conditions. In this work, the first factor is the most relevant to achieve a classification. Levels of Ca, Na, K, and Fe in wines can be influenced by regional variations in fertilizing practice and wine processing. In this case, the low Fisher weight of these features compared to the Fisher weight of Li and Rb indicates that they do not have great relevance as a basis for typification. The content of selected trace elements (Li and Rb) in wine can be used to obtain a highly successful region classification. Use of all available features is unnecessary and undesirable, because the use of variables with no extra discriminating information only introduces noise in the pattern recognition process. Samples from non-Rías Baixas brands can be correctly classified, and this fact allows for the detection of consumer fraud since the probability of classification of a non-Rías Baixas brand as a genuine Rías Baixas brand is nil in practice. However, genuine Rías Baixas wines can be rejected as false.

ACKNOWLEDGMENT

We express our gratitude to Consello Regulador da Denominaci6n de Orixen Rías Baixas for providing wine samples from this producing area. We are grateful to Ana Chouciño (University of Kansas) for valuable observations.

LITERATURE CITED

Armanino, C.; Forina, M.; Castino, M.; Piracci, A.; Ubigli, M. Chemometrical investigation on four red wines from a single cultivar grown in the Piedmont Region. *Analyst* **1990**, *115*, 907–910.

Cruz Ortiz, M. C.; Saez, J. A.; Palacios, J. L. Typification of alcoholic distillates by multivariate techniques using data from chromatographic analyses. *Analyst* **1993**, *118*, 801–805.

Etievant, P.; Pascal, S.; Bouvier, J. C.; Symonds, P.; Bertrand, A. Varietal and geographic classification of French red wines in terms of elements, amino acids and aromatic alcohols. *J. Sci. Food Agric.* **1988**, *45*, 25–41.

Fernández-Pereira, C.; Ortega, J.; Martin, A. Contribution of major and trace metals to the characterization of Spanish wines. *Alimentaria* **1987**, *1*, 39–44.

Forina, M.; Lanteri, S. Data analysis in food chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; Riedel Publishing: Dordrecht, Holland, 1984.

Forina, M.; Leardi, R.; Armanino, C.; Lanteri, S. *PARVUS an extendable package of programs for exploration, classification and correlation*; Elsevier: Amsterdam, 1988.

Herrero, C.; Mèdina, B. Use of some mineral elements in differentiation of Galicia wines. *Connais. Vigne Vin* **1990**, *24*, 147–156.

Kowalski, B. R.; Bender, C. F. Pattern recognition. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.

Kwan, W. O.; Kowalski, B. R. Classification of wines by applying pattern recognition to chemical composition data. *J. Food Sci.* **1978**, *43*, 1320–1323.

Kwan, W. O.; Kowalski, B. R. Pattern recognition analysis of gas chromatographic data. Geographic classification of wines of *Vitis Vinifera cv* Pinot Noir from France and the United States. *J. Agric. Food Chem.* **1980**, *28*, 356–359.

Kwan, W. O.; Kowalski, B. R.; Schogerboe, R. K. Pattern recognition of elemental data. Wines of *Vitis Vinifera cv* Pinot Noir from France and the United States. *J. Agric. Food Chem.* **1979**, *27*, 1321–1326.

Lacasta, F. "*Dossage de quelques mètaux dans les vins par spectrometrie d'absorption atomique*"; Rapport, B. T. A. O.; Station Agronomique-Oenologique de Talence: Bourdeaux, France, 1982.

Latorre, M. J.; Herrero, C.; Mèdina, B. Use of mineral elements to differentiate Galician wines. *J. Int. Sci. Vigne Vin* **1992**, *3*, 185–193.

Maarse, H.; Slump, P.; Tas, A. C.; Schaefer, J. Classification of wines according to type and region based on their composition. *Z. Lebensm. Unters. Forsch.* **1987**, *184*, 198–203.

Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: New York, 1979.

Massart, D. L.; Kaufman, L. Hierarchical clustering methods. In *The interpretation of analytical data by use of cluster analysis*; Wiley: New York, 1983.

Meloun, M.; Militky, J.; Forina, M. Scaling, weighting, transforms. In *Chemometrics for analytical chemistry*; Ellis Horwood: New York, 1992a.

Meloun, M.; Militky, J.; Forina, M. Clustering. In *Chemometrics for analytical chemistry*; Ellis Horwood: New York, 1992b.

OIV. *Recueil des mèthodes internationales d'analyse des vins Office International de Vigne et du Vin*, 5th ed.; OIV Editions: Paris, 1978.

STATGRAPHICS, User's Guide, version 5; STSC: Rockville, MD, 1991.

Vasconcelos, P.; Chaves, H. Classification of elementary wines of *Vitis Vinifera* varieties by pattern recognition of free amino acids profiles. *J. Agric. Food Chem.* **1989**, *37*, 931–937.

Wold, S. Pattern recognition by means of disjoint principal components models. *Pattern Recog.* **1976**, *8*, 127–139.

Wold, S.; Albano, C.; Dun, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johanson, E.; Lindberg, W.; Sjöstrom, M. Multivariate data analysis in chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; Riedel Publishing: Dordrecht, Holland, 1984.

Received for review December 17, 1993. Revised manuscript received March 20, 1994. Accepted April 20, 1994.[⊗]

[⊗] Abstract published in *Advance ACS Abstracts*, May 15, 1994.