Available online at www.sciencedirect.com
ScienceDirect
Food Chemistry 101 (2007) 1585–1597
Food Chemistry
www.elsevier.com/locate/foodchem

Multivariate determination of the geographical origin of wines from four different countries ☆

X. Capron, J. Smeyers-Verbeke *, D.L. Massart ✠

Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 9 November 2005; received in revised form 27 February 2006; accepted 13 April 2006

Abstract

In the framework of the Wine Database European Project, 400 wine samples from four different countries, namely Hungary, Romania, Czech Republic and South Africa, were collected and 63 chemical parameters were analyzed in order to determine the possibility to identify the origin of a wine from its chemical content. The ability of multivariate analysis methods such as classification and regression trees and partial least squares discriminant analysis and its uninformative variable elimination variant to achieve this classification task is investigated and a special attention is given to variable selection. The results observed show that it is possible to obtain excellent classification rates based on the chemical content of only few parameters, such as for instance the isotopic ratios or the concentration in trace elements.

Keywords: Wine discrimination; Classification trees; UVE-PLS; Variable selection

1. Introduction

The European Union has a leading position on the world wine market, accounting for 60% of production and 70% of exports in global terms. Therefore, the wine industry forms an important part of the European Union's economy. However, this market is facing growing competition from imports from countries such as Bulgaria, Romania, the United States, Chile, Argentina, South Africa and Australia. In the past years the official European wine control bodies have been more and more confronted with imported wines which had to be judged as suspicious in terms of their authenticity, e.g. as regards geographical origin.

Within the framework of the Wine Database European Project, a data base containing 400 samples of authentic and commercial wines from Hungary and Czech Republic, two new European member states, Romania, a candidate European member state, and South Africa has been created. Commercial samples are wine samples bought on the market while authentic samples are obtained by micro-vinification of the grapes harvested directly from the vineyard, which guarantees the authenticity of the geographical origin of those samples. The chemical content of each of those samples for 63 parameters (macro and trace elements, isotopic ratios, classical parameters, biogenic amines) has been analyzed. Several parameters concerning the oenological characteristics as well as the geographical origin of each sample were also recorded.

The major aim of the project is to evaluate whether it is possible to determine the country of origin of a wine sample, using its chemical content. Several publications concerning the determination of the geographical origin of

☆ Project Steering Committee: R. Wittkowksi BfR, Germany, P. Brereton CSL, United Kingdom, E. Jamin Eurofins, France, X. Capron VUB, Belgium, C. Guillou JRC, Italy, M. Forina UGOA, Italy, U. Roemisch TU, Berlin, Germany, V. Cotea - UIASI.VPWT.LO, Romania, E. Kocsi NIWQ, Hungary, R. Schoula CTL, Czech Republic.

* Corresponding author. Tel.: +32 2 477 45 12; fax: +32 2 477 47 35.
E-mail address: asmeyers@fabi.vuc.ac.be (J. Smeyers-Verbeke).

✠ Prof. D.L. Massart passed away on December 26, 2005.

0308-8146/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodchem.2006.04.019

wines (see e.g. Christoph et al., 2004; Christoph, Rossmann, & Voerkelius, 2003; Gremaud, Quaile, Piantini, Pfammatter, & Corvi, 2004) using the natural abundance of elements and their isotopes. Here we extend this by also taking into account the wine content in biogenic amines, macro elements and classical parameters.

It was investigated if it is possible to discriminate between the four countries on the basis of the chemical data in the database. Additionally, it was the aim to investigate if discrimination models could be developed with the authentic wines that could be used for the commercial ones. Moreover, this article focuses on three multivariate techniques which have only been applied rarely to authenticate the origin of wine samples, namely the classification and regression trees (CART), the partial least squares-discriminant analysis (PLS-DA) and its uninformative variables elimination (PLS-UVE) variant. This variant eliminates useless, i.e. in this context non-discriminating, variables. To this variant a step was added, which is called PLS-UVE-SEL, the aim of which is to select a small set of variables that is still able to achieve a good discrimination. Therefore, one objective of the project is to obtain efficient models which are economic in terms of the number of required measurements.

2. Theory

A mathematical model must be built that is able to identify from which of the four participating countries a wine sample comes, based on the value of (some of) the 63 chemical parameters determined during the project. In the ideal case, such a model would have the following properties:

- The geographical origin of a sample is always determined correctly, i.e. the correct classification and prediction rates are equal to 100%.
- The model needs as few parameters as possible. In an ideal situation, only one variable would be enough.

2.1. CART

Introduced by Breimann, Friedman, Olshen, and Stone (1984), CART is a non-parametric method, i.e. no assumptions about the distribution of data is required. There have been only a few applications in chemistry (see Caetano, Aires-de-Sousa, Daszykowski, & Vander Heyden, 2005; Put et al., 2003). CART can deal with regression problems, when the response y is continuous, and with classification problems when y is discrete. In this study, our concern is about classification and we therefore focus on this aspect of the method.

CART yields a classification tree by splitting the data into subsets, called nodes, which are more homogeneous (more pure) with respect to the classes than the initial set. The homogeneity of a node can be evaluated by means of several different criteria such as the Gini index, the entropy criterion or the twoing criterion (see Breimann et al. for detailed descriptions of those indices). Although those criteria are very similar in their principle, their different mathematical definitions yield different trees, though their performances are very similar, and hence the Gini index is retained. This index reaches its minimum value when the node contains only objects of the same class, i.e. when the node is pure.

The splitting process starts by the division of the root node, containing all available samples, in two. The procedure is recursive, since the two child nodes obtained from the root node are then treated as parent nodes and split again into two subsets.

The split is done with respect to a cutoff value for the variable that yields the purest child nodes. Depending on whether the variable value of a given sample is lower or higher than the cutoff value, the sample is placed in the left or right child node. Nodes obtained after a split can be either terminal nodes also called leaves, or parent nodes, which are further split by CART. The splitting of nodes continues until terminal nodes are obtained. These are nodes that are considered sufficiently homogeneous, i.e. all samples in the node belong to the same class, or a predefined (small) number of objects in the terminal nodes is reached.

Building the CART model requires three steps:

- First, the data recursive partitioning is carried out until the tree perfectly describes the input data. This tree has a large number of terminal nodes and over-fits the data. This means that the maximal tree obtained in this way describes perfectly the data used to build the tree but that it may have a low predictive ability because it also models the noise present in the data. Hence, it is essential to determine a smaller tree which is a better compromise between the predictive ability and the fit to the data. This determination must be done in the two following distinct steps.
- The last branches of the over-large tree size are successively cut. This procedure, called pruning, determines a sequence of smaller trees.
- Finally, the optimal tree must be found, i.e. it must be evaluated which of the trees obtained by pruning has the best predictive ability for new samples. When evaluating the discrimination models, the final aim of these models must be borne in mind. This is to be able to correctly classify new samples that were not available yet when the model was developed. Thus it is necessary to predict how good the classification of new samples will be. Simply verifying how many of the samples that were used to develop the model are correctly classified often leads to an overly optimistic classification rate. The estimation of the prediction error (i.e. the misclassification rate) can be done either by using an independent validation set or by using cross-validation (CV). The use of an independent data set requires more data and hence k-fold CV is usually the

favorite method used in the development of the model when the number of samples is not very large. In this procedure the data set is split in k subsets, $(k-1)$ subsets being used to develop a CART model while the left out subset is used to test the predictive ability of this model. This process is repeated k times leaving out a different subset each time and hence building k different trees. The final error of prediction is the overall misclassification rate for the trees of each size. Eventually, the optimal tree is the simplest one among those that have a CV error within one standard deviation error of the minimal CV error. During this study, CART model complexity was assessed by 10-fold CV.

When the optimal model has been chosen using the CV procedure, it is further validated by the classification with independent samples. Indeed, although 10-fold CV yields an estimate of the prediction error of the model, it can still give a too optimistic view of the quality of the discrimination model and it is safer to test the performance of the model with independent samples, for instance samples drawn from the available data with the duplex algorithm (Snee, 1977). The discrimination model is then built again with the data that were not drawn and used to predict the classification of the independent samples. This yields what we will call the prediction rate, i.e. the number of independent samples that is correctly classified. We will call re-substitution rate the number of samples used to build the model that are correctly classified. Normally the prediction rate is lower than the re-substitution rate.

2.2. PLS-DA

Regression methods can be applied to discriminant analysis problems by encoding the class membership of a sample as a number. However, with most methods this approach is constrained to the problem of discrimination between two classes encoded as 0 and 1 or -1 and $+1$. Indeed, encoding three different classes with e.g. numbers 1, 2, 3 would introduce an arbitrary ordering of the classes that would suppose that samples belonging to class no. 1 and class no. 2 are more similar to each other than samples from class no. 1 and class no. 3. Hence, if more than two classes are present in the data, more than one model must usually be developed. Among the different regression algorithms available in the literature, PLS (Geladi & Kowalski, 1986; Sjöström, Wold, & Söderström, 1986) is used during this work. The usual PLS algorithm, called PLS1, is applied. It should be noted that another algorithm, PLS2, could be used for more than two classes, but we preferred PLS1 because its properties are better known.

The PLS algorithm does not suffer from collinearity in the data and noise can be filtered out by limiting the number of factors used by the model, which makes it perform better in many cases than a classical Multiple Linear Regression (MLR) model. PLS regression is a method using latent variables, which are linear combinations of the original variables, also called factors.

The principle of the PLS algorithm is the following. The first PLS factor is built in order to maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$. This means that variables that discriminate most between the two classes present in the data are given a higher weight in the construction of the factor. Then the variance explained by this factor is removed from the data, and another factor, which is orthogonal to the first, is obtained from the residuals of $\mathbf{X}$ and $\mathbf{y}$. One latent variable is rarely sufficient and the number of factors yielding the model with best predictive ability must be determined. Indeed, a model built with too many factors captures not only the information related to the discrimination problem but also the noise and particularities of the calibration data. Hence, its predictive ability is not optimal. Cross-validation is the most popular method to optimize the number of factors of the model and to avoid the problem of over-fitting. The cross-validation procedure used for PLS-DA is very similar to the one applied for CART: at each iteration of the procedure, a model is made based on $(k-1)$ subsets and its predictive ability is assessed on the kth subset for all possible complexities. The CV procedure used in this study for PLS is leave-one-out CV (Martens & Naes, 1989), which means that k is equal to the number of samples used to built the model. When each sample has been left out once, the average prediction error of the model as a function of the number of factors is computed and the model with the lowest root mean square error of cross validation (RMSECV) is retained. Once the optimal complexity A_{opt} is determined, the final PLS model using A_{opt} factors can be built.

The prediction of the class membership of a new sample is achieved by means of:

$$\hat{y}_{\text{new}} = \mathbf{x}_{\text{new}}^{\text{T}}\mathbf{b} \tag{1}$$

where $\hat{y}_{\text{new}}$ (1×1) is the predicted value of y for the new sample, $\mathbf{x}_{\text{new}}$ $(p\times 1)$ is a vector containing the measurements of the p original variables and $\mathbf{b}$ $(p\times 1)$ is the vector containing the regression coefficients associated with each of the X variables. Since classes are encoded here as -1 or $+1$, the decision rule is straightforward: if $\hat{y}_{\text{new}}$ is negative then the new sample is associated to the first class and reciprocally if $\hat{y}_{\text{new}}$ is positive then the new sample belongs to the second class.

2.3. PLS-UVE

PLS models always require measuring all the parameters that were used during calibration. This constraint does not match one of the initial requirements, which is that the origin of a sample can be determined with as few parameters as possible. One solution to this problem is to perform variable elimination and this is done in this study using another approach called uninformative variable elimination for PLS (PLS-UVE).

The aim of PLS-UVE (Centner et al., 1996) is to statistically identify which variables do not carry information important for the discrimination problem and to remove them. Indeed only a fraction of the parameters are associated with important regression coefficients. Therefore, variables with low and/or unstable regression coefficients are likely to be uninformative and unstable and can be removed from the model. Hence, a new PLS model using fewer parameters than the original one can be developed. The performances of the PLS-UVE model are better if the variable elimination reduces the noise present in the data, but this elimination may also slightly decrease the predictive power of the model. The PLS-UVE algorithm can be summarized as follows:

- The optimal complexity A of the classical PLS model is determined on the original data **X** and **y** by cross-validation.
- A matrix **R** containing q randomly artificial variables is generated. Those variables have a very small amplitude (10^{-10} for instance) and correspond to noise, i.e. they are known to be uninformative.
- With a leave one out procedure, n different PLS models using A factors are built on the joint matrix [**X**; **R**] and the vector **y**. This yields n different regression vectors **b** with dimensionality $(p + q \times 1)$.
- The mean and standard deviation of each regression coefficient b_i is computed and a stability criterion c_i is estimated:

$$c_i = \text{mean}(b_i)/\text{std}(b_i) \quad (2)$$

- Since the last q regression coefficients are associated with artificial uninformative variables, a cutoff value equal to the maximum absolute value of c for the artificial variables is computed: Threshold = max(abs(c_{artif})).
- All original variables for which $c_i >$ threshold are retained and constitute the new **X** matrix, $\mathbf{X}_{\text{new}}$.
- The final PLS model is built on $\mathbf{X}_{\text{new}}$ and **y**. The complexity of this new model has to be optimized again since the elimination of uninformative variables may lead to a reduced complexity of the model.

2.4. PLS-UVE-SEL

The number of variables that have some discriminating power and are therefore retained by the PLS-UVE method is often rather large. In our study, typically some 20 of the 63 variables are retained. Not all of them may be necessary to achieve a good enough discrimination and the user may prefer to select a smaller set of variables that still yields acceptable performance.

This is the purpose of the selection step. The selection is usually performed stepwise, e.g. by going from 20 variables first to 10, checking how good the result is, if it is good enough then trying only 5, etc.

There are several ways to make the selection, namely:

- Selection of the variables that have the highest b-values. According to Eq. (1) these variables should have the highest influence on y and therefore on the discrimination.
- Selection of the variables with the most stable regression coefficients (c-coefficients).
- Selection based on practicality, e.g. by eliminating first those variables which are analytically the most time consuming to obtain.

It is of course also possible to use more than one criterion during the selection. There is little or no experience about whether selection based on b is to be preferred to selection based on c and vice versa and therefore this point is investigated in detail.

3. Data set

The data base constructed during the first year of the WineDB project contains 393 wine samples from four different countries and is divided into two categories: authentic and commercial wines (see Table 1).

The authentic samples of the European wines are from the 2002 vintage and the commercial samples from 2001. The authentic South African samples were harvested in 2002, the commercial ones in 2003.

For each wine sample many chemical parameters are available, namely the concentrations of certain trace elements, macro elements and biogenic amines, ratios of isotopes and also the measure of so-called classical parameters such as the concentration in glycerol, malic acid, etc... Some computed oenological parameters such as the excess concentration of sodium were added to the list of variables, as well as some rare earth ratios. Eventually 63 parameters were available for the discrimination of the four different countries. The list of variables is summarized in Table 2.

A preliminary study showed that most of the parameters are log normally distributed or at least show a more symmetrical distribution after log transformation. Since some

Table 1
Distribution of data for the four countries

	Red wines	White wines	Total
(a) Authentic samples			
Hungary	15	34	49
Czech Republic	12	37	49
Romania	16	34	50
South Africa	14	36	50
$\sum$	57	141	198
(b) Commercial samples			
Hungary	17	33	50
Czech Republic	16	29	45
Romania	13	37	50
South Africa	13	37	50
$\sum$	59	136	195

Table 2
The list of the 63 variables measured

1	Invert sugar
2	Tartaric acid
3	D-Lactic acid
4	L-Lactic acid
5	Malic acid
6	Glycerol
7	Butanediol
8	Gluconic acid
9	Shikimic acid
10	Methanol
11	Ethylacetate
12	1-Propanol
13	2-Methyl-1-propanol
14	2-Methylbutan-1-ol
15	3-Methylbutan-1-ol
16	Original malic acid
17	Na
18	Mg
19	Si
20	P
21	S
22	Cl
23	K
24	Ca
25	Na_Exc
26	Ethanolamine
27	Putrescine
28	Ethylamine
29	Li
30	B
31	Al
32	Ti
33	V
34	Cr
35	Mn
36	Fe
37	Co
38	Ni
39	Cu
40	Zn
41	As
42	Br
43	Rb
44	Sr
45	Y
46	Cd
47	Cs
48	Ba
49	Pb
50	U
51	La
52	Gd
53	Er
54	Yb
55	Ethanol $(D/H)_1$
56	Ethanol $(D/H)_2$
57	Ethanol $\delta^{13}C$
58	Wine $\delta^{18}O$
59	Gd/La
60	Er/La
61	Yb/La
62	Gd/Er
63	Er/Yb

statistical methods assume the normal distribution of the data, such parameters were log transformed. Additionally, for PLS modeling, the data were autoscaled, i.e. each variable has zero mean and unit variance.

4. Results

4.1. Exploratory analysis by PCA, correlations

Exploratory analysis of both data sets, authentic and commercial, is carried out first (Fig. 1a and b) since it gives a first idea of the complexity of the problem to solve and of the factors that are most important in determining the variance within the data.

The first three PCs describe only 46% and 43% of the total variance present in the authentic and commercial data, respectively. However, it is possible to draw a few conclusions from this PCA analysis. First, the origin of the samples is important and, in fact, the most important factor in the chemical composition of the samples. Further analysis (not shown) indicates that the type (red or white) of wine is the second most important factor. South Africa seems to be very easy to discriminate from the other countries. Indeed, South African samples form a cluster nicely separated from the rest of the data, both for authentic samples (Fig. 1a) and commercial samples (Fig. 1b). The clustering tendency is most visible for the authentic wines. In Fig. 1a, it can be seen that South Africa, Romania and Czechia form three distinct clusters, while Hungary overlaps with Czechia and to a lesser extent with Romania. Therefore we expect discrimination involving Hungarian wines and most of all between Hungarian and Czech wines to be the most difficult one. This is due in part to the heterogeneity of the Hungarian wines. In particular wines from the Tokaj region show different patterns compared to the other samples. For the commercial samples, no clear cluster can be seen and a strong overlap is observed between the three countries. However, Fig. 1b displays only the scores of samples along the first three PCs, which describe not even half of the data variance. As a consequence, even if the discrimination is expected to be more difficult for the commercial samples than for the authentic ones, there may be discriminant information carried by further PCs. Variable selection is also an important aspect of the project and therefore it is essential to have an idea of the correlation between the different parameters. For practical reasons, only the most important correlations between the trace elements are shown in Fig. 2a and b.

4.2. CART

The CART method is applied to authentic and commercial samples separately and the 63 available parameters are input to the method.

4.2.1. Authentic samples

A first CART model is built to discriminate between the four countries of the project. The tree derived from the CART model is represented in Fig. 3a.

This first tree is built on the full data set, without any sample left out for testing purposes. The resulting tree

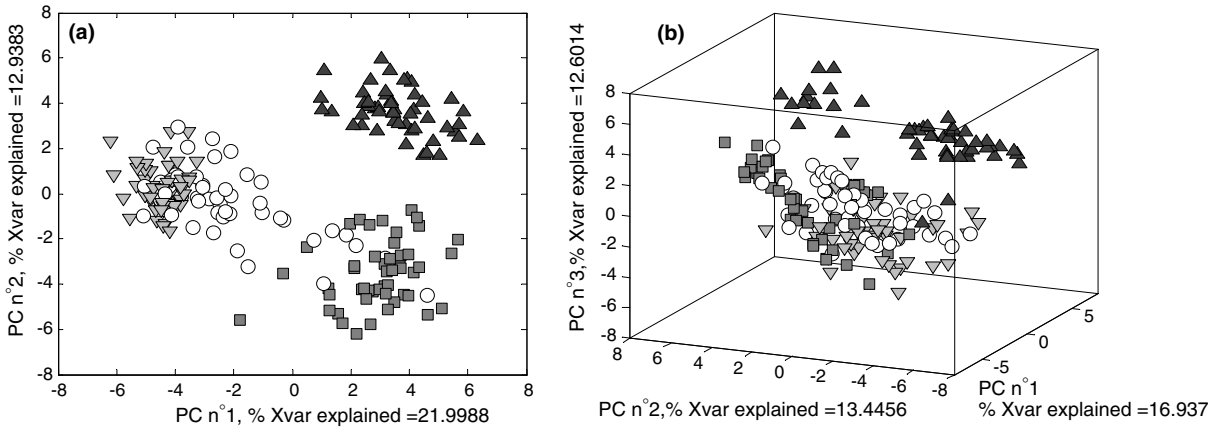

Fig. 1. PCA scores of Hungarian (○), Czech (▽), Romanian (■) and South African (▲) samples for (a) authentic wines; (b) commercial wines.

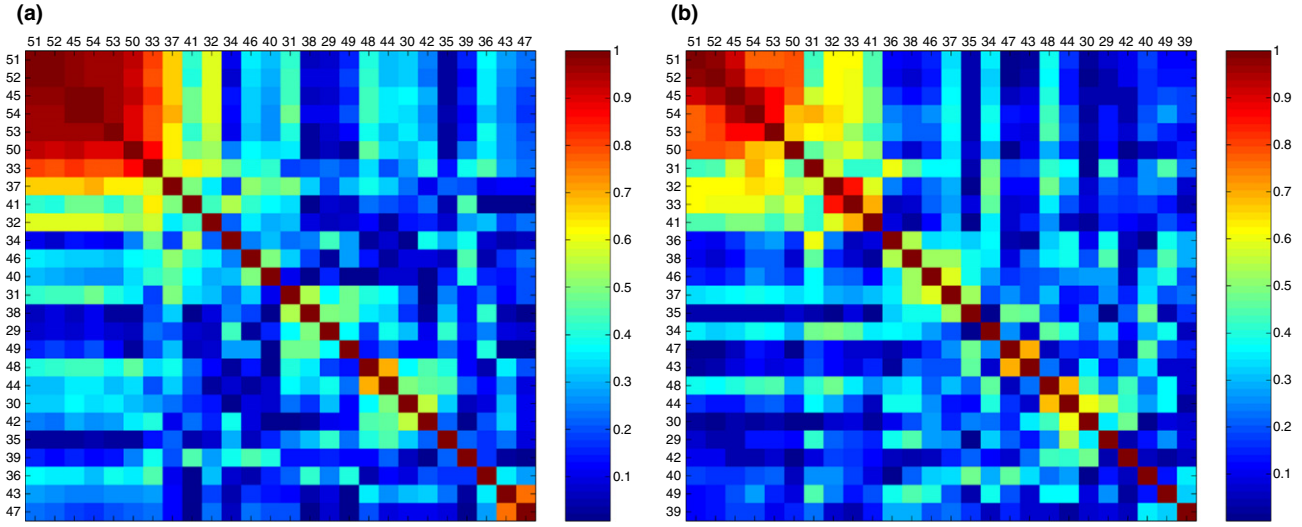

Fig. 2. Correlation map of the trace elements for (a) authentic samples; (b) commercial samples. Numbers correspond to the index of variables used in Table 2.

clearly confirms the first conclusion from PCA, i.e. South African samples are very easy to discriminate from the other authentic wines. Indeed, the correct re-substitution rate achieved by CART for South Africa is 100% and uses only the value of one isotopic ratio, namely ethanol $(D/H)_1$, the ratio between deuterium and hydrogen in the CH_3 group of ethanol. The boxplot of this parameter (see Fig. 4a) shows that South African wines are indeed characterized by a higher value of this isotopic ratio.

Moreover, the CART method makes it possible to find which variables would lead to a similar discrimination as the one achieved with the chosen parameter. In this case, ethanol $(D/H)_1$ could be replaced by ethanol $(D/H)_2$, the deuterium to hydrogen ratio in the CH_2 group of ethanol, or wine $\delta^{18}O$, the ratio between ^{18}O and ^{16}O in the water of the wine (see boxplots in Fig. 4b and c). This underlines the importance of the isotopic measurements to discriminate South African wines from the other samples. Further research, when more samples are available will indicate which of the three, or which combination of the three parameters is to be preferred.

This first tree also shows that discrimination between Hungary, Romania and Czech Republic is not as straightforward. As expected, some Hungarian samples are wrongly classified as Romanian or Czech samples. As a consequence, it was decided to build a second tree in order to discriminate only between Hungarian, Romanian and Czech authentic wines. For validation purposes, this reduced data set is split into two independent sets by means of the Duplex algorithm (Snee, 1977). The calibration set contains 99 samples, i.e. 33 samples from each country, and the test set contains 16 Hungarian, 16 Czech and 17 Romanian samples for a total of 49 wine samples. Samples in the independent test set are not used during the elaboration of the model and are used to assess the error of prediction of the classification tree. The CART tree obtained from this second model is represented in Fig. 3b. It uses three chemical parameters to discriminate between the

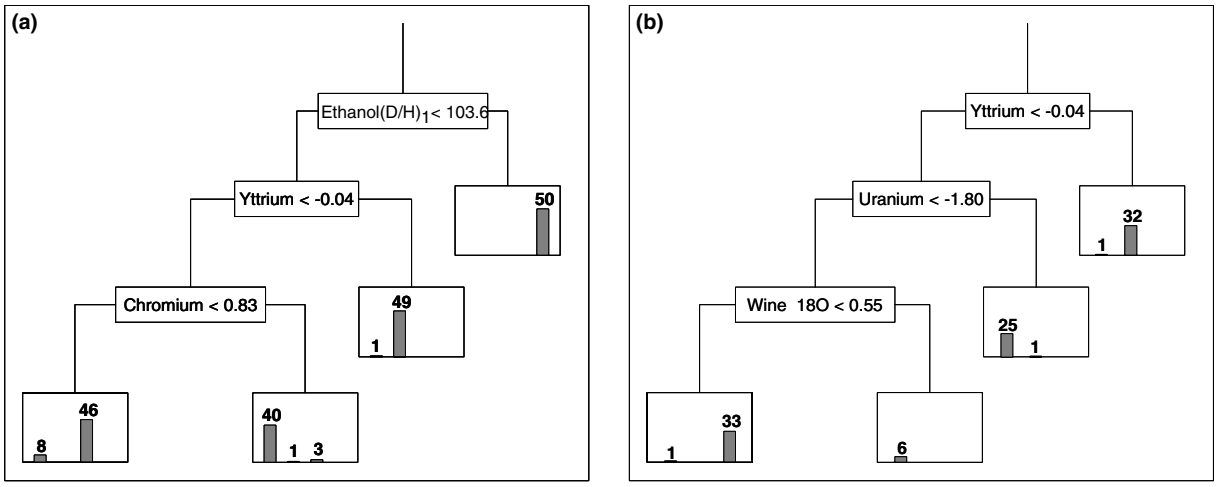

Fig. 3. Classification trees obtained from CART for (a) authentic samples from the four countries; (b) authentic samples from Hungary, Romania and Czech Republic. Bars in boxes stand for Hungarian, Romanian, Czech and South African samples, respectively.

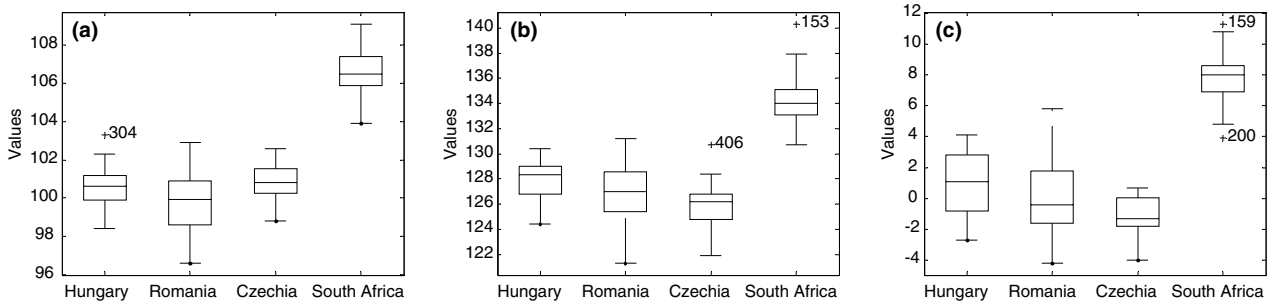

Fig. 4. Boxplots of (a) ethanol $(D/H)_1$; (b) ethanol $(D/H)_2$; (c) wine $\delta^{18}O$ for authentic samples.

different wine samples, two trace elements (Y, U) and one isotopic ratio (wine $\delta^{18}O$). As expected from the PCA analysis, Romania is separated first. This is done on the basis of the Y concentration. The two other variables are needed to separate the Hungarian and Czech samples. In Fig. 3a, Cr is used instead of U to discriminate between Hungarian and Czech wines. However, CART tells us that those two parameters have very close discriminating power, which explains the change observed. Only three samples of the 99 are misclassified, which is equivalent to a 3% re-substitution error. The prediction rate is slightly worse, 8% of samples being misclassified. Those results are satisfying taking into account the small number of variables used to develop the model.

4.2.2. Commercial samples

The first CART model built on commercial data again tries to discriminate between the four countries of the project (Fig. 5a). As expected, South African samples are very easy to discriminate from wine samples from eastern European countries.

The measurement of ethanol $(D/H)_1$, the same isotopic ratio as for authentic samples, is sufficient to carry out this discrimination. As for the authentic samples ethanol $(D/H)_2$ and wine $\delta^{18}O$ can replace ethanol $(D/H)_1$ to discriminate South Africa from the other countries. Moreover, the splitting values of ethanol $(D/H)_1$ for authentic and commercial samples are almost similar, 103.6 and 102.8, respectively, which makes it possible to identify all South African wine samples by comparison of this isotopic ratio with a reference value chosen equal to 103, for instance. It can be concluded that for this discrimination the authentic wines are a good model for the commercial ones.

A second CART model is built for the commercial samples from the three European countries. The data is split with the Duplex algorithm in a calibration set containing 99 samples (33 samples from each country) and a test set consisting of 17 Hungarian, 17 Czech and 12 Romanian samples. The obtained classification tree (Fig. 5b) uses four parameters, namely the two of the first tree (B and wine $\delta^{18}O$) and additionally the trace element Pb and one biogenic amine (ethanolamine). The latter is particularly important to discriminate between Hungary and the two other countries. Seven training samples and seven test samples are misclassified when their class membership is predicted by the model, which yields a re-substitution rate of 93% and the prediction rate is 85%. The prediction rate is relatively low but the CART tree uses only four variables, which can explain the relatively poor predictive ability of the model.

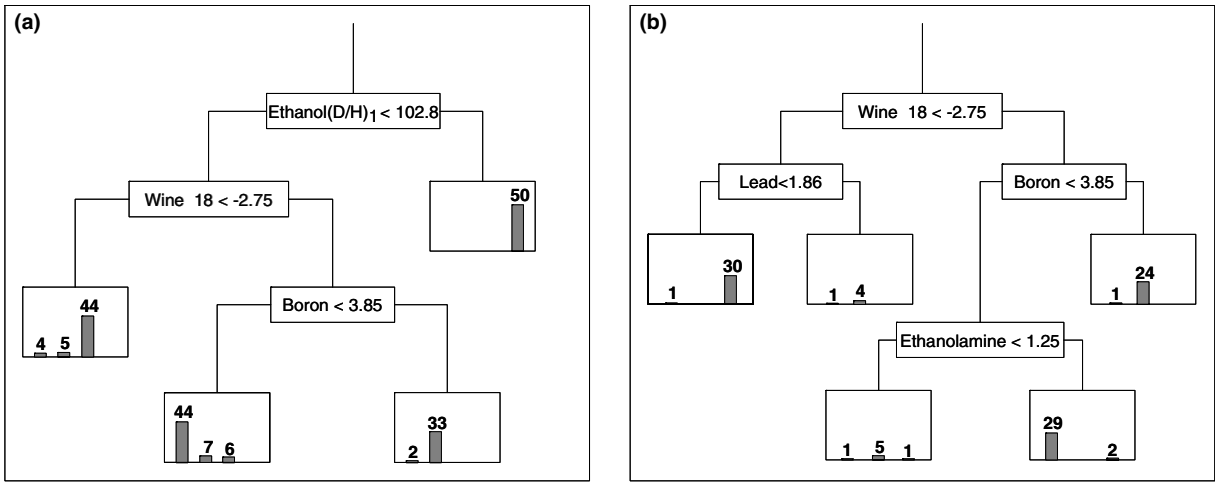

Fig. 5. Classification trees obtained from CART for (a) commercial samples from the four countries; (b) commercial samples from Hungary, Romania and Czech Republic. Bars in boxes stand for Hungarian, Romanian, Czech and South African samples, respectively.

4.2.3. Other CART models

We tried to build CART models to discriminate between white wines of the four countries. There are too few red wine samples to build specific trees for those samples. The models obtained have either the same performance as the model built on red and white wines together in the case of authentic samples, or behave worse in the case of commercial samples where the prediction error increases to 25%. This may be surprising since the classification problem should be simplified. However fewer samples are available for the elaboration of the model, which can explain the lower predictive ability of the tree.

It was also checked if the CART models built with authentic samples from Hungary, Romania and Czech Republic can be applied to the discrimination of the commercial samples, i.e. if the model built with authentic samples can discriminate between commercial samples. However, the predictive abilities of the trees are found to be unacceptable. Indeed, the correct prediction rates in this situation do not exceed 36%. It is concluded that the original hypothesis that a discrimination model for the authentic wines might serve to develop a model for commercial wines is not verified here.

For practical reasons, it can be better to work with a limited number of types of parameters, e.g. only isotopic ratios or isotopic ratios and elements. Modern analytical devices measure for instance the concentration of all trace and macro elements at the same time, which is very economic. Therefore, CART was also applied to the situations where only isotopic ratios or only trace and macro element measurements are available but the performances observed were less satisfying.

4.3. 3/PLS-DA, UVE-PLS and PLS-UVE-SEL

Since South African samples can be discriminated from the three others in a univariate way, only the European countries are considered further. The constraints inherent to PLS-DA make it necessary to develop several different models since it is not possible to discriminate more than two classes at the same time. Two alternatives are possible: build pair-wise discriminant models (i.e. Hungary vs. Romania, Hungary vs. Czechia and Romania vs. Czechia, once for authentic samples and once for commercial samples) or build one vs. all others discriminant models. The latter is preferred here because the aim of this study is to determine if a given wine sample originates from where it claims to. If a wine sample is supposed to be Hungarian, the model should discriminate between Hungary vs. {Czechia + Romania}. Therefore, three different PLS-DA models are developed: Hungary vs. {Czech Republic + Romania}, Czech Republic vs. {Hungary + Romania}, Romania vs. {Hungary + Czech Republic}, once for authentic data and once for commercial data. Classical PLS and UVE-PLS models are compared in terms of performance.

4.3.1. Authentic samples

Discrimination Romania-{Hungary + Czech Republic}: The optimal PLS-DA model uses two PLS factors and the correct classification and prediction rates are 99% and 100%, respectively. The PLS scores (Fig. 6) show that there is no overlap between the two classes, though one Romanian sample is very close to the border, and illustrate the excellent discrimination results obtained.

The first PLS component separates the Romanian from all Czech and most Hungarian samples, the second PLS component completes the separation. The Hungarian samples discriminated by the second PLS component are mainly Tokaj samples and the second PLS component is therefore due to the heterogeneity within the Hungarian samples. The UVE-PLS approach retains 25 variables, mainly trace elements (cf. Table 3) and the PLS model built with these variables has the same predictive ability

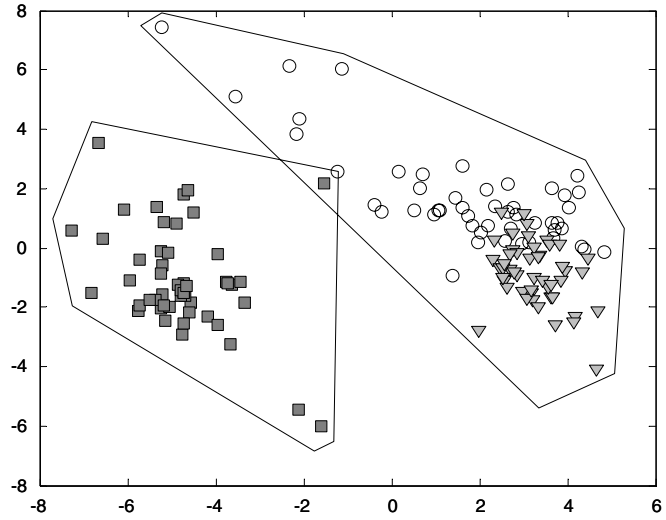

Fig. 6. PLS scores of authentic Romanian (■), Hungarian (○) and Czech samples (▽).

as the PLS-DA model using the whole set of parameters. As the number of parameters retained by UVE-PLS is usually high, those models will not be described in further details in the remaining of this article. The next question is whether the number of parameters can be further reduced by the application of PLS-UVE-SEL. Let us first consider the selection based on the stability coefficients c.

According to c the most stable variables are La (+ all other rare earths and Y), U and V. Because of the very high correlation between the rare earths (RE) and Y, we count these as one variable and will do so throughout the further analysis of the data. The selection of these variables seems logical, when comparing it with the CART results and the box plots, since Y and U are among the best discriminating variables and this is also true for V. A PLS-DA model built with Y, U, V and the RE only uses two factors and yields re-substitution and prediction rates of 96% and 99%, respectively. The selection on c therefore works well. The samples that are not well classified by re-substitution or prediction are the five Hungarian Tokaj samples that are not discriminated along the first PLS component in Fig. 6. When only the RE + Y are retained, a two factor PLS model results with a re-substitution error of still only 4% and a prediction error of now 4% (i.e. one Tokaj sample more than when U and V are included).

Table 3
List of the parameters retained by UVE-PLS and their stability coefficients

Variables PLS-UVE	Stability
Ethanolamine	35.8
Tartaric acid	37.9
La	179.8
Er	165.6
Y	152.8
Yb	143.1
Gd	164
Original malic acid	37
Co	85.6
V	106.5
Si	39.3
Yb/La	82.5
Gd/La	74.6
U	130.6
Cu	22.8
Er/La	79.7
Sr	27.1
Li	24.2
Er/Yb	68.2
B	30.3
S	46
Gd/Er	58.4
Na	28.9
Gluconic acid	23.3
Na_Exc	30.4

In Table 3 the variables are ranked according to their b-coefficients. According to b the most important variables are ethanolamine, tartaric acid and La (+ all other RE and Y). Looking at the box plots (not shown here) the selection based on ranking according to b seems much less logical than that based on c. Indeed, ethanolamine appears to have no discrimination power at all. However, when a model is built with ethanolamine, tartaric acid and the RE + Y, a very good result is obtained since the re-substitution error is 2% and the prediction error 0%. This seems surprising but can be explained as follows. Variables with a high b-value and a low c-value, such as ethanolamine have a high std dev b, since $c = b/\text{std dev } b$ (see Eq. (2)). This indicates that ethanolamine and tartaric acid are both important for only a few samples. They are responsible for the discrimination of the Hungarian samples that are not separated along the first PLS factor (and most of which were not discriminated by the models selected on the basis of c). By going back to the original data, it is found that the ethanolamine values for the Tokaj region are indeed lower than for the Romanian samples.

The selection on the basis of b for two variables is not satisfying since the model built only on ethanolamine and tartaric acid yields unacceptable re-substitution and prediction errors of 26% and 30%, respectively. Those results could be expected since those two parameters are only interesting to discriminate the Tokaj wine. This highlights the fact that the selection based on b is not automatically the best one. In fact, the best model obtained for these data uses two parameters selected with c (U and V) and two selected with b (ethanolamine and tartaric acid). This model requires two factors and gives re-substitution and prediction errors of 0% and 1%, respectively.

Discrimination Czech Republic-{Hungary + Romania}: The best PLS-DA model requires two factors and yields re-substitution and prediction rates of 98% and 100%, respectively.

As expected from the PCA and CART results, the discrimination between Hungary and the Czech Republic is

the most difficult one. While there is no overlap between the Czech and the Romanian samples, there is a slight overlap between Hungarian and Czech samples in the PLS space (see Fig. 7). However, the PLS-DA model is still able to deliver excellent predictions.

PLS-UVE allows to decrease the number of required variables from 63 to 35. Then two PLS-DA models are built with the five most important variables according to *b* on the one hand and with the five most stable parameters on the other. The former of these two models uses two PLS factors and the correct re-substitution and prediction rates are 97% and 98%, respectively. The PLS model built with the five most stable variables, i.e. Y + La + Gd, U, V, Cr and 3-methylbutan-1-ol gives correct re-substitution and prediction rates of 97% and 94%. The importance of V is not expected from the CART tree but can be explained by its strong correlation with Y, the RE and U which are important discriminating parameters. In this situation, the selection based on *b* is better, though it is difficult to say if the differences observed are significant or not. The best model that could be found uses five variables, namely U, Cr, Yb, 3-methylbutan-1-ol and wine $\delta^{18}O$ and the re-substitution and prediction rates are 99% and 98%, respectively, when the model uses two factors.

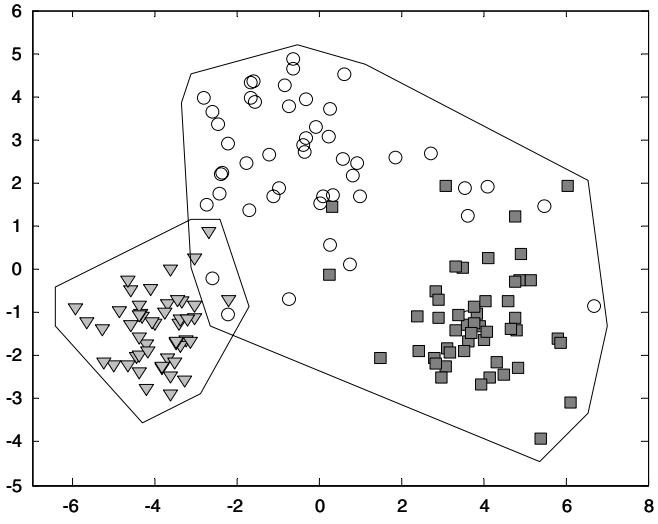

Fig. 7. PLS scores of authentic Czech (▽), Romanian (■) and Hungarian (○) samples.

Discrimination Hungary-{Czech Republic + Romania}: The optimal PLS-DA model requires four factors. The correct re-substitution and prediction rates are both equal to 98%. Two Romanian samples have extreme characteristics (see Fig. 8a, samples no 394 and 351) and it is decided to discard them from the data set. The new PLS model (Fig. 8b) built without those samples requires only two PLS factors and the correct re-substitution and prediction rates are equal to 97% and 98%, respectively. The first PLS factor discriminates the Hungarian and the Romanian wines and the second the Hungarian and the Czech samples.

The PLS-UVE algorithm retains 21 parameters out of 63. The model built with the five most discriminating parameters according to *b* gives unsatisfactory performances. The correct re-substitution and prediction rates are only 90% and 86%, respectively. However, the five most stable variables (Cr, 3-methylbutan-1-ol, ethylacetate, Mn, ratio Er/Yb) yield re-substitution and prediction rates equal to 95% and 94%, which is acceptable considering the low number of parameters used and the difficulty of the discrimination.

Notice that none of the variables important according to CART are retained here. This might be due to the fact that the PLS factors try to find a compromise between the separation of Hungary from Czechia on the one hand and of Hungary from Romania on the other, so that unexpected parameters are used. The classification results could not

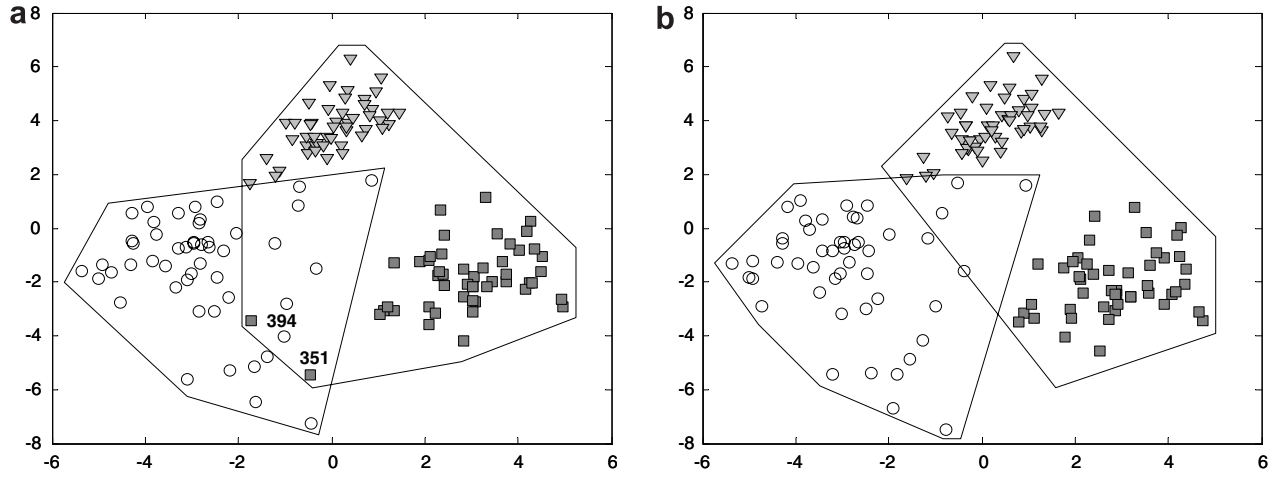

Fig. 8. (a) PLS scores of Hungarian (○), Romanian (■) and Czech (▽) authentic samples along factor 1 and factor 2; (b) PLS scores from the model calibrated without Romanian samples no. 394 and no. 351.

be improved by considering separate discrimination models Hungary/Romania and Hungary/Czechia.

4.3.2. Commercial samples

Discrimination Romania-{Hungary + Czech Republic}: A two factors PLS-DA model has a very satisfying discriminating power. The correct re-substitution and prediction rates are equal to 99% and 100%, respectively. The PLS scores (Fig. 9) show that this discrimination is straightforward to perform.

Twenty-one parameters are retained by UVE-PLS. In order to further reduce the number of parameters necessary to perform the discrimination, two different models are built. For the model built with the five most discriminating parameters according to *b* (ethanolamine, Si, Cd, Zn and Sr) two PLS factors are necessary and the correct re-substitution and prediction rates are 98% and 96%, respectively. The model built with the five variables selected according to the value of *c* (Sr, Cd, B, ethanolamine and ethanol $(D/H)_1$) has a very similar predictive ability. The correct re-substitution and prediction rates are 97% and 96%, respectively. The best model is built with ethanolamine, Sr, Cd, wine $\delta^{18}O$ and La. The correct re-substitution and prediction rates are 99% and 100%, respectively. The presence of La is surprising since it is not one of the most important parameters according to either *b* or *c*.

Discrimination Czech Republic-{Hungary + Romania}: The optimal PLS-DA model is built with two factors and has correct re-substitution and prediction rates of 99% and 100%, respectively. According to the UVE-PLS approach, 16 of the 63 available variables are important (mainly isotopic ratios and trace elements). As for the other discriminations, two models built with only five variables are developed. The first model uses the five parameters with the highest *b* values (Ti, Zn, Pb, wine $\delta^{18}O$, L-lactic acid). This model requires two factors and the correct re-substitution and prediction rates are 97% and 90%, respectively. The second model uses two factors and is built with the five most stable parameters (ethanol $(D/H)_1$, ethanol $(D/H)_2$, ethanol $\delta^{13}C$, wine $\delta^{18}O$, Zn) and the resulting re-substitution and prediction rates are 95% and 83%, respectively.

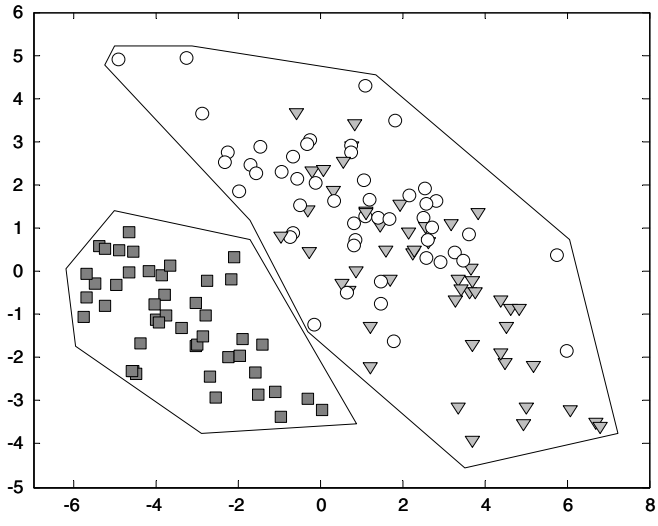

Fig. 9. PLS scores of commercial Romanian (■), Hungarian (○) and Czech samples (▽).

Given the importance of the isotopic values to achieve this discrimination, a PLS-DA model is built only on the four isotopic ratios available. This model uses two factors and its correct re-substitution and prediction rates are equal to 93% and 85%, respectively. This model yields classifications results very similar to the model built on the five most stable variables since four of those variables are isotopic ratios, the last parameter being Zn.

The best PLS-DA model developed is built with four parameters, i.e. Ti, Cu, wine $\delta^{18}O$ and invert sugar. This model has correct re-substitution and prediction rates of 97% and 94%, respectively.

Discrimination Hungary-{Czech Republic + Romania}: A PLS-DA model using three factors has a very good discriminating power, the correct re-substitution and prediction rates are 99% and 100%, respectively. Of the 63 initial parameters, 18 are retained by the PLS-UVE algorithm. The five parameters with highest *c* value, namely ethanolamine, putrescine, B, Sr and wine $\delta^{18}O$, are used to develop a PLS model with two factors. This model yields correct re-substitution and prediction rates equal to 88% and 90%, respectively. Ethanolamine, B and wine $\delta^{18}O$ are also found in the CART tree. Sr mainly discriminates Hungary and Romania and replaces Pb and the box plots show that putrescine indeed has some discriminating power. Another model using two PLS factors built with the five parameters with highest *b* (ethanolamine, Ca, Ti, P, B) gives correct re-substitution and prediction rates equal to 93% and 85%, respectively. The best PLS-DA model obtained is built with eight parameters: ethanolamine, wine $\delta^{18}O$, P, Ca, Li, U, Sr and B. This model uses four factors and the correct re-substitution and prediction rates are 95% and 96%, respectively. When fewer parameters are included, the number of misclassified samples increases immediately, as shown by the models built with only five parameters.

5. Conclusions

The analyses clearly show that South African wines are very easy to discriminate from the wines of Eastern Europe. Indeed, the measurement of ethanol $(D/H)_1$ is sufficient to tell if a wine sample is from South Africa or not, whether this sample is authentic or commercial. For the commercial wines this conclusion must be viewed with some caution since the wines are not from the same vintage year and it is known that isotopic ratios can change very much from year to year. However, as the same discrimination is obtained for the authentic wines which were harvested in the same year it seems probable that isotopic

ratios are also of value in the discrimination of South African commercial wines.

The discrimination of Hungarian wine samples from Czech and Romanian ones is the most difficult one, while the discrimination between Czech and Romanian wines is quite straightforward. This seems to indicate again that the difficulty of the discrimination is related to the geographical distance between the countries of interest.

The discriminating power of the different models is very satisfying. CART models, though they are univariate and use very few parameters to determine the country of origin of a sample, yield correct prediction rates, especially for authentic samples, where three parameters are enough to correctly identify a wine sample 92% of the time. For commercial samples, CART does not give such good results but the correct prediction rate of 85% is still satisfying taking into account that only four parameters and very simple decision rules are necessary to achieve this performance.

With PLS discriminating models using the whole set of 63 parameters, it is possible to achieve perfect discrimination almost all the time. However, the best compromise between economy and quality of discrimination, particularly for commercial samples, is to use the UVE-PLS approach followed by UVE-PLS-SEL.

For a minimal cost in terms of predictive power, usually less than 5%, it is possible to achieve very good discrimination of the different wine samples. For instance, it is possible to discriminate correctly commercial samples from Hungary, Czech Republic and Romania 97% of the time measuring only six variables on the average.

It is not possible to conclude whether the selection of parameters on the basis of regression coefficients *b* is better than the one based on the stability coefficients c. It depends on the discrimination considered. If both *b* and *c* are high, then this parameter should be kept and a variable with both low *b* and *c* values should be discarded. When b is high and *c* is low, it is likely that this parameter is only important to discriminate few samples, and depending on the importance of those samples, it should be discarded or not. Finally, parameters with a low *b* and high *c* are stable but with low discriminating power and hence could be eliminated.

It must be underlined that the models described in this article are built for the first vintage year of the project. Since wines are depending on the vintage, it is probable that models presented here must be updated in order to deliver the same quality of prediction.

The discriminant models developed here for authentic samples are not able to classify correctly commercial wine samples. This implies that models built for authentic samples cannot be used to discriminate commercial samples. However, this may be due to the fact that the authentic and commercial samples are not from the same vintage year.

This research was undertaken with the aim of evaluating and optimizing two discriminating methods, namely CART and PLS, with variable selection. It will be completed at a later date when more samples are available by comparing results with those obtained by linear discriminant analysis and related methods (Vandev & Römisch, 2004) and SIMCA (Wold et al., 1983) and multivariate range modeling.

Acknowledgements

The authors acknowledge the contributions of the European commission for the financial support of this work, which was carried out in the framework of the specific research and technological development program "Competitive and Sustainable Growth" (Contract G6RD-CT-2001-00676). The authors are solely responsible for the content of this research article and the European Community is not responsible for any use that might be made of the data appearing therein.

Finally, the authors thank all the partners of the European WineDB project and especially the ones who collected the different wine samples, performed the micro-vinification of the authentic samples and did the analytical measurements of all the parameters necessary to carry out this study.

References

Breimann, L., Friedman, J. H., Olshen, R. A., & Stone, C. G. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

Caetano, S., Aires-de-Sousa, J., Daszykowski, M., & Vander Heyden, Y. (2005). Prediction of enantioselectivity using chirality codes and classification and regression trees. *Analytica Chimica Acta, 544*, 315–326.

Centner, V., Massart, D. L., de Noord, O. E., de Jong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry, 68*, 3851–3858.

Christoph, N., Baratossy, G., Kubanovic, V., Kozina, B., Rossmann, A., Schlicht, C., et al. (2004). Possibilities and limitations of wine authentication using stable isotope ratio analysis and traceability. Part 2: wines from Hungary, Croatia and other European countries. *Mitteilungen Klosterneuberg, 54*, 155–169.

Christoph, N., Rossmann, A., & Voerkelius, S. (2003). Possibilities and limitations of wine authentication using stable isotope and meteorological data, data banks and statistical tests. Part 1: wines from Franconia and Lake Constance 1992 to 2001. *Mitteilungen Klosterneuberg, 53*, 23–40.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*, 1–17.

Gremaud, G., Quaile, S., Piantini, U., Pfammatter, E., & Corvi, C. (2004). Characterization of Swiss vineyards using isotopic data in combination with trace elements and classical parameters. *European Food Research and Technology, 219*(1), 97–104.

Martens, H., & Naes, T. (1989). Multivariate calibration. Chicester, UK: Wiley & Sons.

Put, R., Perrin, C., Questier, F., Coomans, D., Massart, D. L., & Vander Heyden, Y. (2003). Classification and regression tree analysis for molecular descriptor selection and retention prediction in

chromatographic quantitative structure–retention relationship studies. *Journal of Chromatography A, 988*, 261–276.

Sjöström, M., Wold, S., & Söderström, B. (1986). PLS discriminant plots. In *Proceedings of PARC in Practice*. North-Holland: Elsevier Science Publishers B.V.

Snee, R. D. (1977). Validation of regression models: method and examples. *Technometrics, 19*, 415–428.

Vandev, D., & Römisch, U. (2004). Comparing several methods of discriminant analysis on the case of wine data. *Pliska Stud. Math. Bulg., 16*, 299–308.

Wold, S., Albano, C., Dunn, K. W. J., III, Helberg, S., Johansson, E., & Sjostrom, M. (1983). *Pattern recognition: finding and using regularities in multivariate data. Food Research and Data Analysis*. Barking: Applied Science Publishing, pp. 147.