Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions

O. Galtier, O. Abbas, y. Le Dréau, C. Rebufa, J. Kister, J. Artaud, N. Dupuy

Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions

O. Galtier, O. Abbas, Y. Le Dréau, C. Rebufa, J. Kister, J. Artaud, N. Dupuy*

ISM², UMR 6263, équipe AD²EM, groupe Systèmes Chimiques Complexes, case 451, Université d'Aix-Marseille III, 13397 Marseille cedex 20, France

Keywords:
Classification
Discriminant analysis
PLS-DA
SIMCA
Variable selection
Infrared
Crude petroleum oils
Virgin olive oils

A B S T R A C T

This study compares results obtained with several chemometric methods: SIMCA, PLS2-DA, PLS2-DA with SIMCA, and PLS1-DA in two infrared spectroscopic applications. The results were optimized by selecting spectral ranges containing discriminant information. In the first application, mid-infrared spectra of crude petroleum oils were classified according to their geographical origins. In the second application, near-infrared spectra of French virgin olive oils were classified in five registered designations of origins (RDOs). The PLS-DA discrimination was better than SIMCA in classification performance for both applications. In both cases, the PLS1-DA classifications give 100% good results. The encountered difficulties with SIMCA analyses were explained by the criteria of spectral variance. As a matter of fact, when the ratio between inter-spectral variance and intra-spectral variance was close to the F_c (Fisher criterion) threshold, SIMCA analysis gave poor results. The discrimination power of the variable range selection procedure was estimated from the number of correctly classified samples.

1. Introduction

Pure pattern recognition techniques are oriented in discriminated way among different groups of samples and operate by dividing the hyperspace in as many regions as the number of groups. So, if a sample is represented in the region of space corresponding to a particular category, it is classified as belonging to that category. In this case, each sample is always assigned to one and only one group [1]. These methods include Linear Discriminant Analysis (LDA) [2] and Partial Least-Squares Discriminant Analysis (PLS-DA) [3]. The principle of PLS-DA consists in a classical PLS regression where the response variable is binary and expresses a class membership. Therefore, PLS-DA does not allow attributing a sample to other groups than the ones first defined. As a consequence, all measured variables play the same role with respect to the class assignment. Actually, PLS latent variables are built to find a proper compromise between two purposes: describing the set of explanatory variables and predicting the response ones. A PLS-DA classification should well benefit from such a property in the direction of building typologies with an intrinsic prediction power.

Another group of class-modeling techniques represents a different approach to pattern recognition, as it focuses on modeling the analogies among the elements of a class rather than on discriminating among the different categories. In these methods, each category is modeled separately. The objects in agreement with the model are considered as a member of the class, while objects not in agreement are rejected as non-members. When more than one class is modeled, three different situations can be encountered: each sample can be assigned to a single category, or represented by several categories or not be included in any category. In comparison with pure pattern recognition techniques, class-modeling tools offer at least two main advantages: it is in principle possible to recover samples which are not represented in any of the examined categories and which, as a consequence, can be either simply outlying observations or members of a new class not considered during the modeling stage. Moreover, as each category is modeled separately, any additional class can be added without recalculating the already existing class models. The most commonly used chemometric class-modeling technique is SIMCA (Soft Independent Modeling of Class Analogy) [4,5].

The range of study by supervised pattern recognition techniques is wide. Some recent reviews about applications of these chemometric techniques have been published: general reviews on

* Corresponding author at: Université Paul Cezanne, Faculté des Sciences et Techniques de Saint Jérôme, Avenue escadrille Normandie Niemen, 13397 Marseille, France.
E-mail address: nathalie.dupuy@univ-cezanne.fr (N. Dupuy).

geographical origin of foods [6,7] or reviews on source rock origin [8], wines [9], and honeys [10].

There have been many reports in the literature comparing the performance of different pattern recognition techniques on spectroscopic data [11–14]. The increase of pattern recognition techniques and pattern recognition applications led us to search the best method to be used.

The aim of this study was to compare the results obtained with SIMCA, PLS2-DA, PLS2-DA with SIMCA, and PLS1-DA in two applications, after optimization on the basis of spectral variance analysis. The first application concerns a classification of crude petroleum oils according to their geographical origins using mid-infrared (MIR) spectroscopy. The second application is about the classification of French virgin olive oils in five registered designation of origin (RDOs) by near-infrared (NIR) spectroscopy.

2. Experimental

2.1. Samples

2.1.1. Crude petroleum oils

Crude petroleum oils (36 samples), from different fields, were analyzed by MIR spectroscopy to identify their geographical origins: Algeria (ALG), $n = 11$; South America (S.A.), $n = 7$; Equator (EQU), $n = 11$; and Venezuela (VEN), $n = 7$. Five MIR spectra were recorded for each sample and models were performed on these five spectra per sample. Replicates have been collected following. Spectra have not been recording origin by origin to not bias chemometric treatments. Hence, a total of 180 spectra which were divided in two subgroups: the calibration ($n = 23 \times 5 = 115$) set samples which were chosen to take into account all possible variations because of natural variations; and the prediction set samples ($n = 13 \times 5 = 65$) which were randomly selected for each geographic origin. All the replicates of the same sample were used in the same set.

2.1.2. Virgin olive oils

Commercial virgin olive oils (317 samples) were obtained from the French Inter-Professional Olive Oil Association (AFIDOL, Aix-en-Provence, France) and from Service Commun des Laboratoires du ministère des finances français (SCL, Marseille, France). Samples were obtained from four successive crops (from 2003/2004 to 2006/2007). They came from five French RDOs: Aix-en-Provence (AP) $n = 97$, Haute-Provence (HP) $n = 46$, Nice (NI) $n = 46$, Nyons (NY) $n = 41$, and Vallée des Baux de Provence (VB) $n = 87$. Samples were analyzed by NIR spectroscopy. Spectra have not been recorded origin by origin, not to bias chemometric treatments.

The calibration set was made up of 225 samples chosen in order to take into account all possible variations because of natural variations among fatty acid and triacylglycerol compositions. The determination of fatty acid and triacylglycerol rates in these samples have been described in previous work [7]; fifty samples having the highest and the lowest fatty acid and triacylglycerol concentrations were chosen in calibration. The other samples were randomly selected. The prediction set was made up of 92 samples which were not selected in the calibration set. The years of harvest are not used as a criterion and all of them could be found in calibration and prediction sets.

2.2. Spectroscopic techniques

2.2.1. Mid-infrared spectroscopy

MIR spectra of each crude petroleum oils were recorded five times from 3400 to 700 cm^{-1}, with 4 cm^{-1} resolution and 100 scans on a Nicolet Avatar spectrometer equipped with a DTGS detector, an Ever-Glo source and a KBr/Germanium beam splitter. The spectrometer was placed in an air-conditioned room (21 °C). Samples were deposited without preparation on a single bounce attenuated total reflection (ATR) cell provided with a diamond crystal. Air was taken as reference for the background spectrum before each sample. Between each spectrum, the ATR plate was cleaned in situ by scrubbing with ethanol solution and dried. Cleanliness was verified by a comparison between a new background spectrum and the previous background spectrum. The recorded spectra have been normalized after correction of the baseline by the instrument software OMNIC 4.1b (Thermo Nicolet). Each spectrum is constituted of 1402 points.

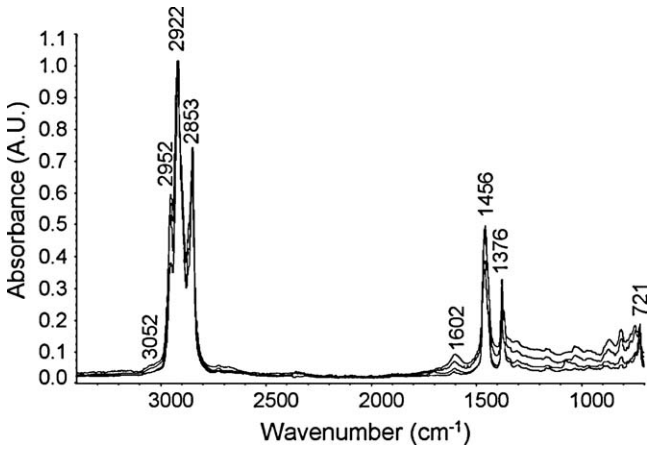

Fig. 1. MIR normalized spectra of crude petroleum oils. Absorption bands: $\nu_{=C-H}$: 3052 cm^{-1}; ν_{as} CH_3: 2952 cm^{-1}; ν_{as} CH_2: 2922 cm^{-1}; ν_s CH_2: 2853 cm^{-1}; $\nu_{C=C}$: 1602 cm^{-1}; δ_{as} C–H in CH_3 and CH_2 groups:1456 cm^{-1}; δ_s C–H in CH_3 group: 1376 cm^{-1}; δCH_2 in $-(CH_2)_{n,\, n>3}$: 721 cm^{-1}.

2.2.2. Near-infrared spectroscopy

NIR spectra of each virgin olive oil were recorded with a Nicolet Antaris spectrometer interfaced to a personal computer using the software result integration 2.1 Thermo Nicolet 2.1. Virgin olive oil samples were filled into a 2 mm pathlength quartz cell directly sampled from the bottle without any chemical treatment. Spectra were recorded between 4500 and 10,000 cm^{-1} at 4 cm^{-1} resolution by co-adding 10 scans using double sided interferograms and an empty cell as a reference. The recorded spectra have been normalized by the UNSCRAMBLER software before chemometric applications. Each spectrum is constituted of 2853 points.

2.3. Unsupervised pattern recognition

Principal Component Analysis (PCA) is an unsupervised pattern recognition and it is often the first step of exploratory data analysis to detect groups in the measured data. PCA models the directions of maximum variations in a data set by projecting as a swarm of points in a space defined by principal components (PCs). PCs describe, in decreasing order, the higher variations among the objects, and because they are calculated to be orthogonal to another one, each PC can be interpreted independently. That permits an overview of the data structure by revealing relationships between the objects as well as the detection of deviating objects. To find these sources of variations, the original data matrix is decomposed into the object space, the variable space, and the error matrix. The error matrix represents the variations not explained by the previously extracted PCs and is dependent on the problem definition [15,16]. The PCA algorithm is used with mean centered data.

2.4. Supervised pattern recognition

2.4.1. Soft Independent Modeling of Class Analogy Classification (SIMCA)

SIMCA is the most used of the class-modeling techniques. The SIMCA classification is a method based on disjoint PCA modeling

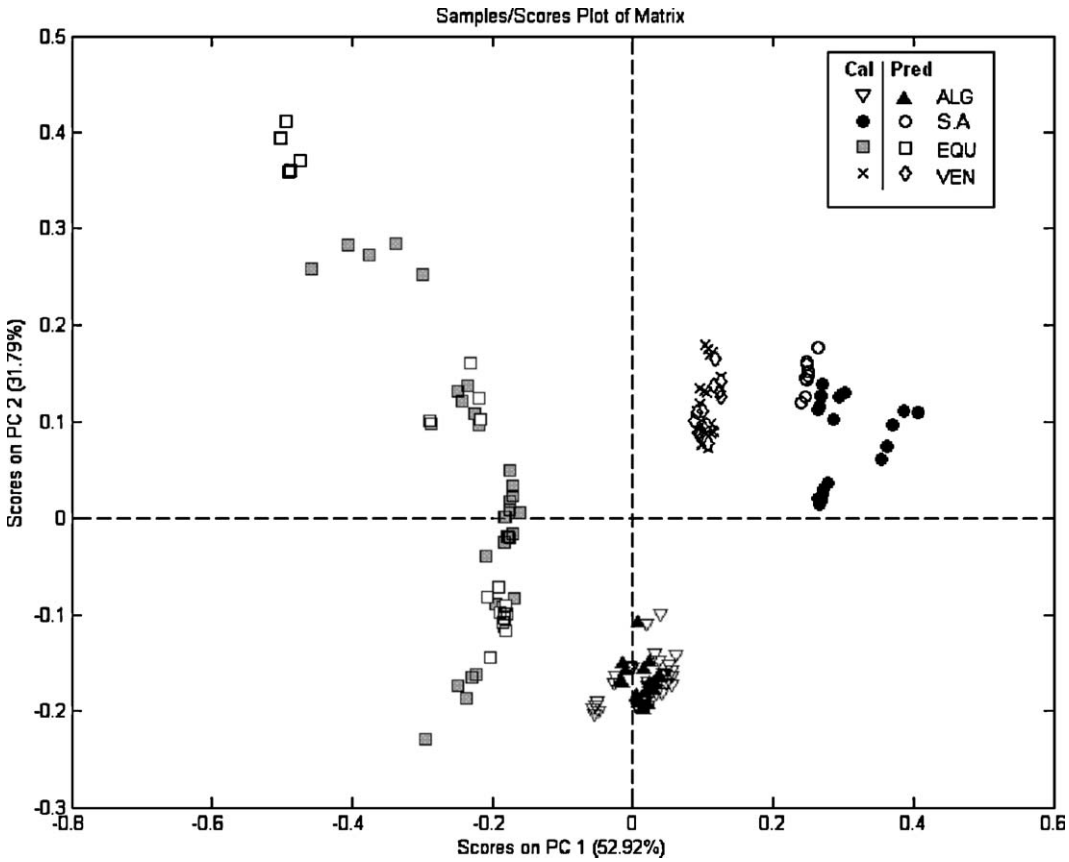

Fig. 2. PCA of the crude petroleum oil spectra on PC1 (53%) and PC2 (32%). ALG: Algeria; S.A.: South America; EQU: Equator; VEN: Venezuela.

realized for each class in the calibration set. Unknown samples are then compared to the class models and assigned to classes according to their analogy with the calibration samples. A new sample will be recognized as a member of a class if it is enough similar to the other members; else it will be rejected. Each class is modeled using separate PCA models. A model distance limit S_{max} is used for classifying new samples and S_{max} is calculated for the class model m as follows (Eq. (1)):

$$S_{max}(m) = S_0(m)\sqrt{F_c} \quad (1)$$

where S_0 is the average distance within the model, and F_c (Fisher criterion) is the critical value provided by the Fisher–Snedecor tables. The F_c value depends on the percentage of risk, generally set to 5% [17]. Class membership is defined at a significance level of 2.5% of S_{max}. Mean centering is applied before modeling.

2.4.2. Partial least squares regression (PLS)

PLS [18,19] was initially built for quantitative analysis, but now it is also used for pattern recognition. This supervised analysis is based on the relation between spectral intensity and sample characteristics [20]. Interference and overlapping of the spectral information may be overcome using powerful multicomponent analysis such as PLS regression. The ability of this algorithm is to mathematically correlate spectral data to a property matrix (relative rate or geographical origin) [21]. Mean centering is applied before modeling. The number of latent variables selected for the PLS model was obtained by cross validation on the calibration set.

When several dependent data are available for calibration, two approaches can be used in PLS regression: either properties are calibrated for one at a time (PLS1), or properties are calibrated at once (PLS2). In PLS1 model, the Y response consists of a single variable. When there is more than one Y response a separated model must be constructed for each Y response. In PLS2 model, responses are multivariate. PLS1 and PLS2 models provide different prediction set and PLS2 regression give better results than PLS1 regression only if Y variables are strongly correlated [1,11,22]. In the other case, PLS1 models are generally more robust [22,23].

PLS regression can be adapted for pattern recognition, giving rise to the PLS-DA method. PLS-DA is performed using an exclusive binary coding. During the calibration process, the PLS-DA method is trained to compute the "membership values", one for each class; the sample is then assigned to one class when the value is above a specific prediction threshold [24]. This method, adapted from PLS1 or PLS2 regressions, uses M spectral variables as predictors and q variables (0 or 1) as variables response [25–28].

The predicted origins seldom lead to a binary result not exactly equal 0 or 1 but to a result near 0 or 1, which is justified by the natural variability of the sample constituents. In front of the difficulty of calibrating and predicting origin with binary variables, it is necessary to discriminate the results between the initial values 0 or 1. Samples with values lower than 0.5 and higher than 1.5 were identified as outside the defined origin and samples with values between 0.5 and 1.5 were identified as belonging to the defined origin.

For PLS1-DA, one regression for each class has been build. For PLS2-DA, all the classes are included in one regression. For PLS2-DA-SIMCA, the SIMCA classification is performed on the PLS2-DA scores. The number of latent variables selected for the PLS model was obtained by cross validation on the calibration set.

2.5. Data processing

2.5.1. Comparison of the methods

The percentage of correct classification (%CC) is the criterion used to compare classification results obtained with chemometric

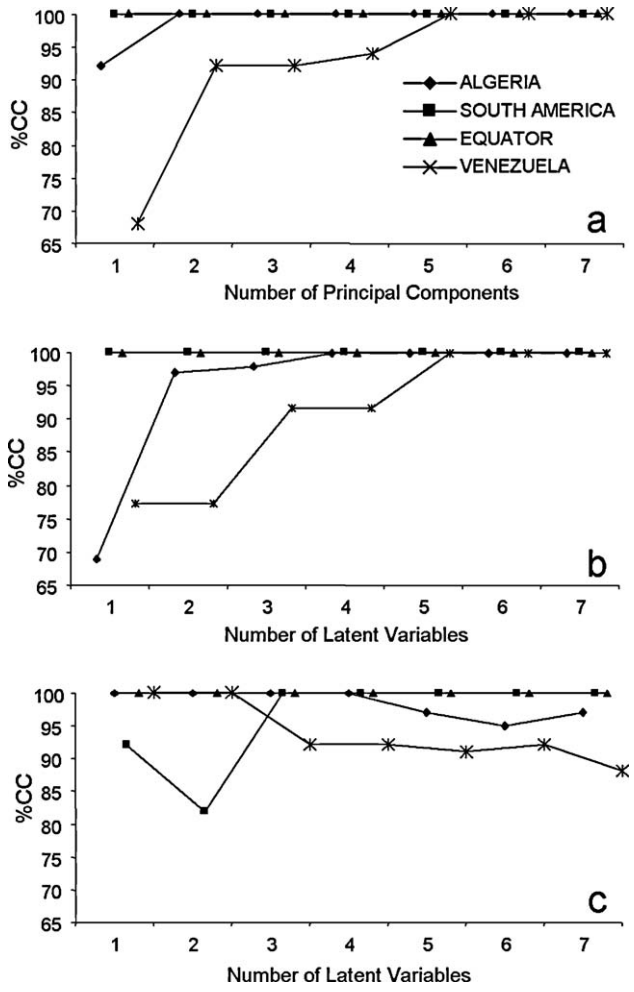

Fig. 3. Percentage of correct classification (%CC) for each petroleum oil origin as function of: (a) the number of principal component used in the SIMCA model; (b) the number of latent variable used in the PLS2-DA model; (c) the number of latent variable used in the PLS1-DA model.

methods:

$$\%CC = \frac{N_c}{N_c + N_{ic}} \times 100 \quad (2)$$

where N_c is the number of correct classifications and N_{ic} is the number of incorrect classifications [29].

2.5.2. Selection of variables with variance method

The total sample variance (S^2) of a sample set is explained by the intra-spectral variance (IntraSP) and by the inter-spectral variance (InterSP). IntraSP represents the variance of all spectra into one class (j spectra) calculated at each wavenumber calculated as follows (Eq. (3)):

$$\text{IntraSP} = \frac{\sum_{i=1}^{j}(a_{ix} - \overline{a_x})^2}{j-1} \quad (3)$$

where a_{ix} is the absorbance of the spectrum i at the wavenumber x and a_x is the mean absorbance of all the j spectra into the class considered at the wavenumber x.

InterSP represents variance of all spectra (n spectra) into the sample set as follows (Eq. (4)):

$$\text{InterSP} = \frac{\sum_{k=1}^{n}(A_{kx} - \overline{A_x})^2}{n-1} \quad (4)$$

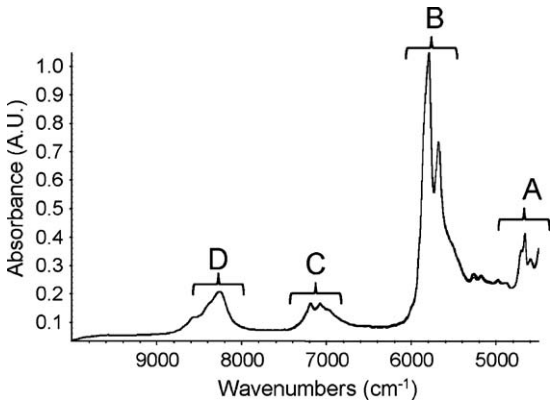

Fig. 4. NIR normalized spectra of virgin olive oil samples. (A) (4500–4800 cm^{-1}) combination of CH stretching vibrations with other vibrational modes; (B) (5300–6100 cm^{-1}) first overtone of CH_2 stretching vibrations (methyl and methylene groups groups); (C) (6700–7450 and 7900–9000 cm^{-1}) combination of CH stretching vibrations, and (D) (cm^{-1}) second overtone of CH stretching vibrations (D: methyl and methylene groups groups).

where A_{kx} is the absorbance of the spectrum k at the wavenumber x and A_x is the mean absorbance of all the n mean spectra of each class considered at the wavenumber x.

Both intra- and extra-spectral variance have been calculated using the Bessel's correction, $n-1$, where n is the number of spectra. InterSP and IntraSP are estimators of the distribution of spectral variances. The Fisher–Snedecor test [30] enables to determine whether a variance is significantly higher than another one by the calculation of the Fisher–Snedecor variable (F) according to Eq. (5):

$$F = \frac{\text{InterSP}}{\text{IntraSP}} \quad (5)$$

IntraSP is an estimator of the intra-group variance σ_1^2, InterSP is an estimator of the inter-group variance σ_2^2.

The null hypothesis H_0 ($\sigma_1^2 = \sigma_2^2$) is tested against the alternative hypothesis H_1 ($\sigma_1^2 = \sigma_2^2$). If the ratio $(j-1)\text{IntraSP}/\sigma_1^2$ follows a χ^2_{j-1} law and if the ratio $(n-1)\text{IntraSP}/\sigma_2^2$ follows a χ^2_{n-1} law, so that, under the null hypothesis $\sigma_1^2 = \sigma_2^2$, F is distributed as $[\chi^2_{n-1}/j-1]/[\chi^2_{m-1}/n-1]$ which does not depend on the common variance σ^2 of the two normal distributions, and can therefore be used as a test statistic.

The value F is compared to the Fisher criterion (F_c). F_c is the critical value provided by the Fisher–Snedecor tables obtained by freedom degree number of interSP and by freedom degree number of IntraSP. F_c depends on the percentage of risk, generally set to 5% [17]. If $F < F_c$, the null hypothesis H_0 is verified and the classification is impossible. If $F > F_c$, the null hypothesis H_0 is not verified and the classification is possible. In this case, IntraSP is significantly lower than the InterSP. Thus, the corresponding spectral range will be selected for model construction. Discrimination power of variable selection procedure is estimated from the number of correctly classified samples.

2.5.3. Software

The chemometric applications were performed by the UNSCRAMBLER software version 9.6 from CAMO (Computer Aided Modelling, Trondheim, Norway).

3. Results

3.1. Classification of crude petroleum oils according to their geographical origins

Crude petroleum oils extracted from different petroleum fields [Algeria (ALG), South America (S.A.), Equator (EQU), and Venezuela

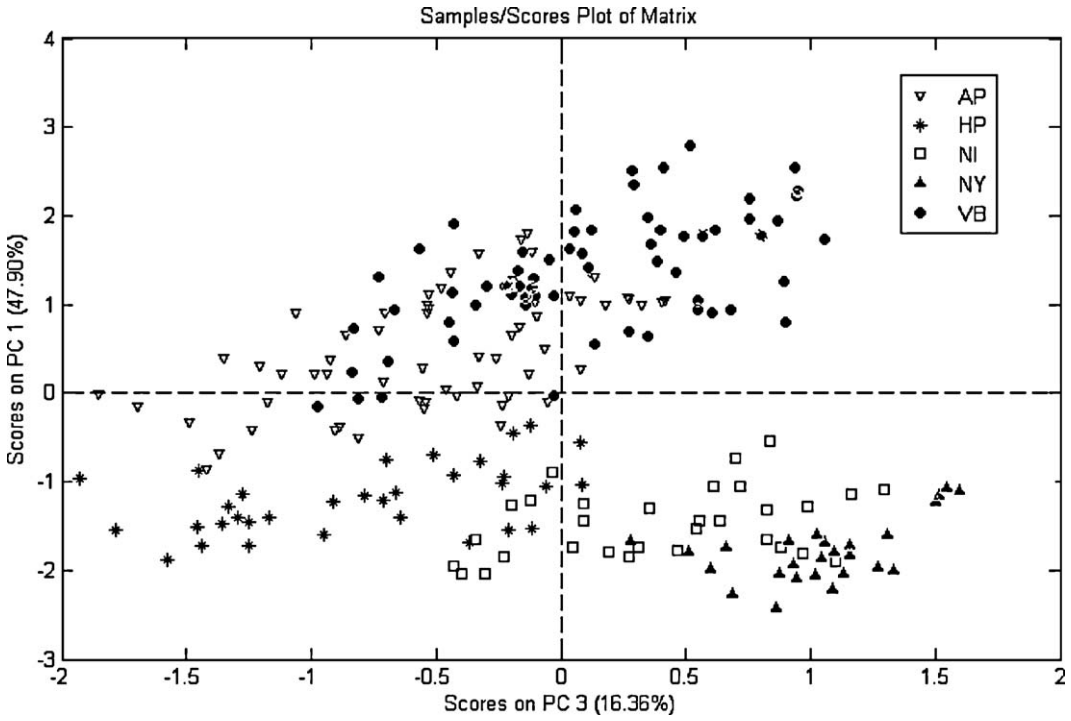

Fig. 5. PCA of the virgin olive oil spectra on PC1 (48%) and PC3 (17%). AP: Aix-en-Provence; HP: Haute-Provence; NI: Nice; NY: Nyons; VB: Vallée des Baux.

Table 1
Percentage of well classified petroleum oils for each statistical treatment in the 700–3400 cm^{-1} spectral range.

Origins	SIMCA		PLS2-DA		PLS2-DA-SIMCA		PLS1-DA	
	%CC	PC	%CC	LV	%CC	LV-PC	%CC	LV
Algerian (ALG)	100	2	100	5	100	4–1	100	1
South American (S.A.)	100	1	100	5	100	4–1	100	3
Equator (EQU)	100	1	100	5	100	4–1	100	1
Venezuelan (VEN)	100	5	100	5	100	4–1	100	1

%CC: correct classification percentage; PC: principal component; LV: latent variable.

(VEN)] were analyzed by MIR spectroscopy. In spite of their different locations, these samples have similar MIR spectra (Fig. 1) and display characteristic bands of aliphatic hydrocarbons: $\nu_{as}CH_3$: 2952 cm^{-1}, $\nu_{as}CH_2$: 2922 cm^{-1}, ν_sCH_2: 2853 cm^{-1}, δ_{as}C–H in CH_3 and CH_2 groups: 1456 cm^{-1}, δ_sC–H in CH_3 group: 1376 cm^{-1}, δCH_2 in $-(CH_2)_n$ ($n > 3$): 721 cm^{-1}. Absorption bands describing aromatic compounds occur at 3052 cm^{-1} ($\nu_{=C-H}$), 1602 cm^{-1} ($\nu_{C=C}$) and in the range 900–746 cm^{-1} (γ_{C-H}) characteristic of the number of adjacent hydrogen atoms on the aromatic ring.

Fig. 2 shows the PCA performed on the full MIR spectra (1402 variables) of crude petroleum oils that constitute four perfectly distinct groups. Prediction samples are highlighted in the score plot projection. The score plots are projected in the PC1-PC2 plane. These PCs explain respectively 53% and 32% of the spectral variance. Three groups are narrowed, and another group (EQU) is farthest. In the light of this PCA overview (85% of the spectral variance is explained with only 2 PCs), a good classification of crude petroleum oils was expected. In our previous work [8], the first principal component was attributed to the aliphatic part of the oil and the second one to oxidised and aromatic compounds.

The samples belonging to ALG, S.A. and VEN groups are very closed to each other unlike the EQU samples which have a large dispersion. This could be explained by an American Petroleum Industry (API) degree and chemical compositions more disparate in the EQU group than in the other ones [31].

The data were treated by different chemometric methods in order to predict their geographic origins. Table 1 shows the best classification results obtained in prediction by each method to predict crude petroleum oil groups according to their geographical origin (ALG, S.A., EQU and VEN). These results were obtained on normalized spectra; spectral treatments as Multiplicative Signal Correction (MSC) or Standard Normal Variate (SNV) have not improved the results. The results of each classification model were 100% satisfactory with all methods. The number of principal components (PCs) and latent variables (LVs) were compared according to the percentage of correct classification. The PLS2-DA method used more LVs than PLS1-DA. The number of LVs in PLS1-DA is lower than the number of PCs used in SIMCA analysis except for S.A. group. Model built on SIMCA analysis of PLS2-DA results was good only with one principal component. Whatever the model, the number of variables (PCs or LVs) allowing to obtain good results is smaller than 5 and almost the same for each.

Fig. 3a shows the percentage of correct classification (%CC) as a function of the number of PCs used in the SIMCA model for each origin. When one principal component is used for all the models, 100% of good classifications are obtained for two origins (EQU and S.A.), 92% of good classification is obtained for ALG and 67% for VEN. The increase of PCs number conduced to 100% of good classification for all the origins. The VEN model obtained with 2 PCs provided to 92% of good classified samples, that represents only 4 misclassified spectra of the 65 analyzed. One hypothesis is that to get good results, it is necessary to have more PCs for models for near groups (ALG and VEN, according to the projection on PC1, Fig. 2) than for other groups that are more distant.

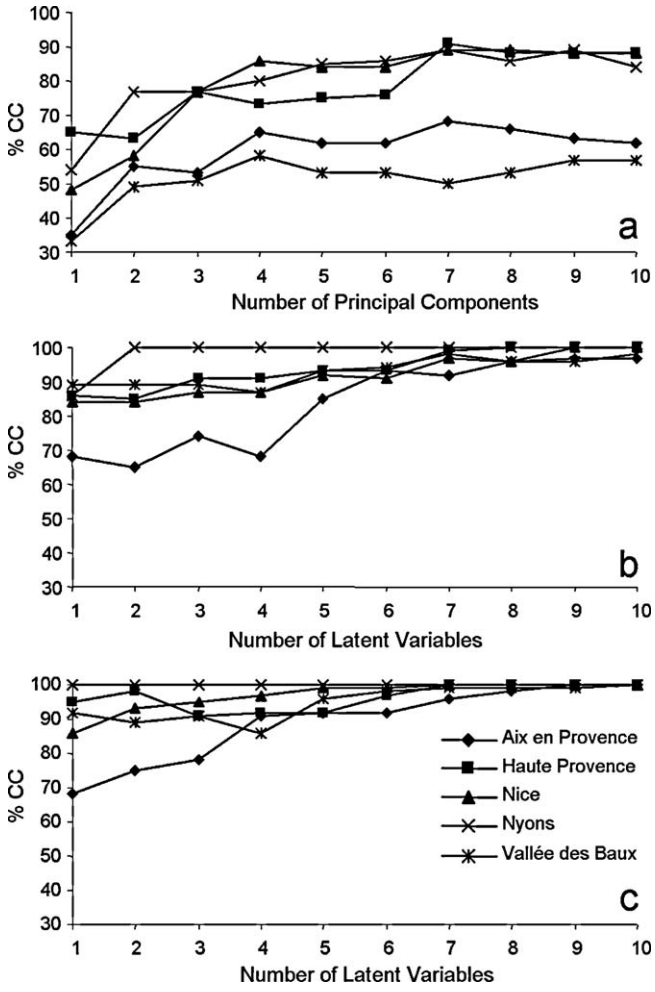

Fig. 6. Percentage of correct classification (%CC) for each virgin olive oil origin as function of: (a) the number of principal component used in the SIMCA model; (b) the number of latent variable used in the PLS2-DA model; (c) the number of latent variable used in the PLS1-DA model.

Fig. 3b shows the %CC as a function of the number of latent variables (LVs) used in the PLS2-DA model. The best model contained 5 LVs even if only 1 LV was necessary for S.A. and EQU origins to well predict all samples, because in PLS2 the same number of LVs must be chosen for all the origins.

Fig. 3c shows the %CC as a function of the number of LVs used in the PLS1-DA model. The number of LVs necessary to obtain the best model for PLS1-DA method was lower than the number of LVs used in PLS2-DA method, and comparable to the number of PCs used in SIMCA analysis. When the number of LVs for the PLS1-DA models increases the results of prediction are altered because of over fitting.

In this case, the variable selection was not necessary because all the petroleum crude oils are correctly classified.

3.2. Classification of virgin olive oils according to their geographical origins

NIR spectra obtained for all virgin olive oil samples seem to be similar (Fig. 4). Band assessments were realized according to the literature [32,33]. Bands A (4500–4800 cm^{-1}) are attributed to combination of CH stretching vibrations with other vibrational modes, bands B (5300–6100 cm^{-1}) are attributed to first overtone of CH_2 stretching vibrations (methyl and methylene groups), bands C (6700–7450 cm^{-1}) are attributed to combination of CH stretching vibrations, and bands D (7900–9000 cm^{-1}) are attributed to second overtone of CH stretching vibrations (D: methyl and methylene groups).

Fig. 5 shows the PCA performed on the full NIR spectra (2853 variables) of the virgin olive oils, which constitute five groups with high overlapping because of the very close RDO origins of the samples. The score plot which allow obtaining the best RDO group separation, are projected in the plane PC1, PC3 and explain 48% and 16% of the spectral variance. Even though the other PCs explain 35% of the spectral variance, plots of PC1 vs. PC2 (19%), PC2 vs. PC3, etc., do not allow a better group separation. The difficulty of obtaining five perfectly distinct groups comes from some similarities of the compositions of these virgin olive oil RDOs. For instance, Aglandau is one of principal cultivars in AP, HP and VB RDOs and Salonenque is the second principal cultivar in AP and VB RDOs with, however, different ratios. NI and NY RDOs are mono-varietal oils constituted respectively by Cailletier and Tanche cultivars.

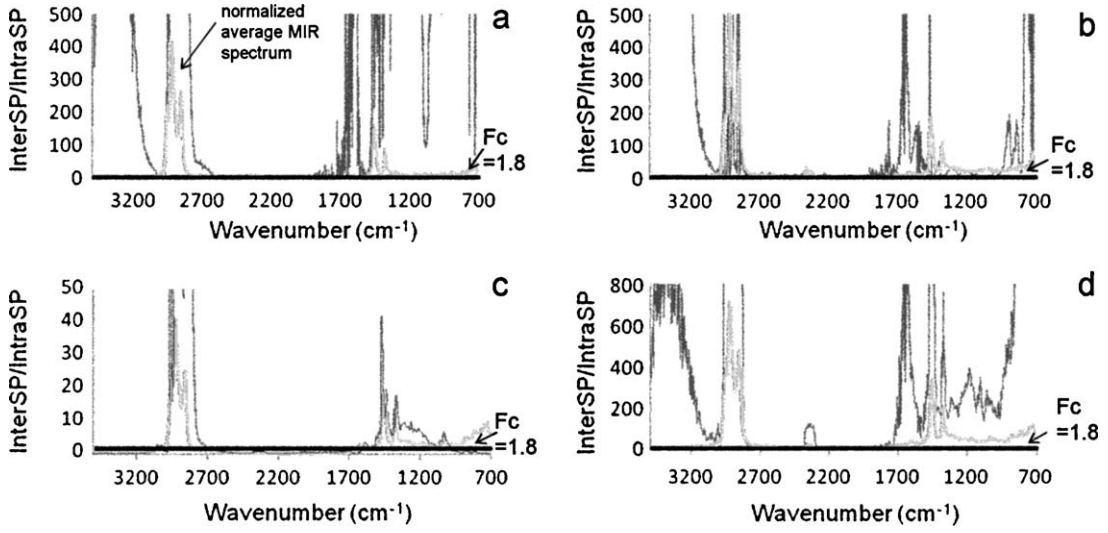

Fig. 7. Ratio of InterSP/IntraSP of each class of petroleum oils. (a) Algerian; (b) South American; (c) Equator; (d) Venezuelan and (e) superposition with the mean spectra of each class and F_c.

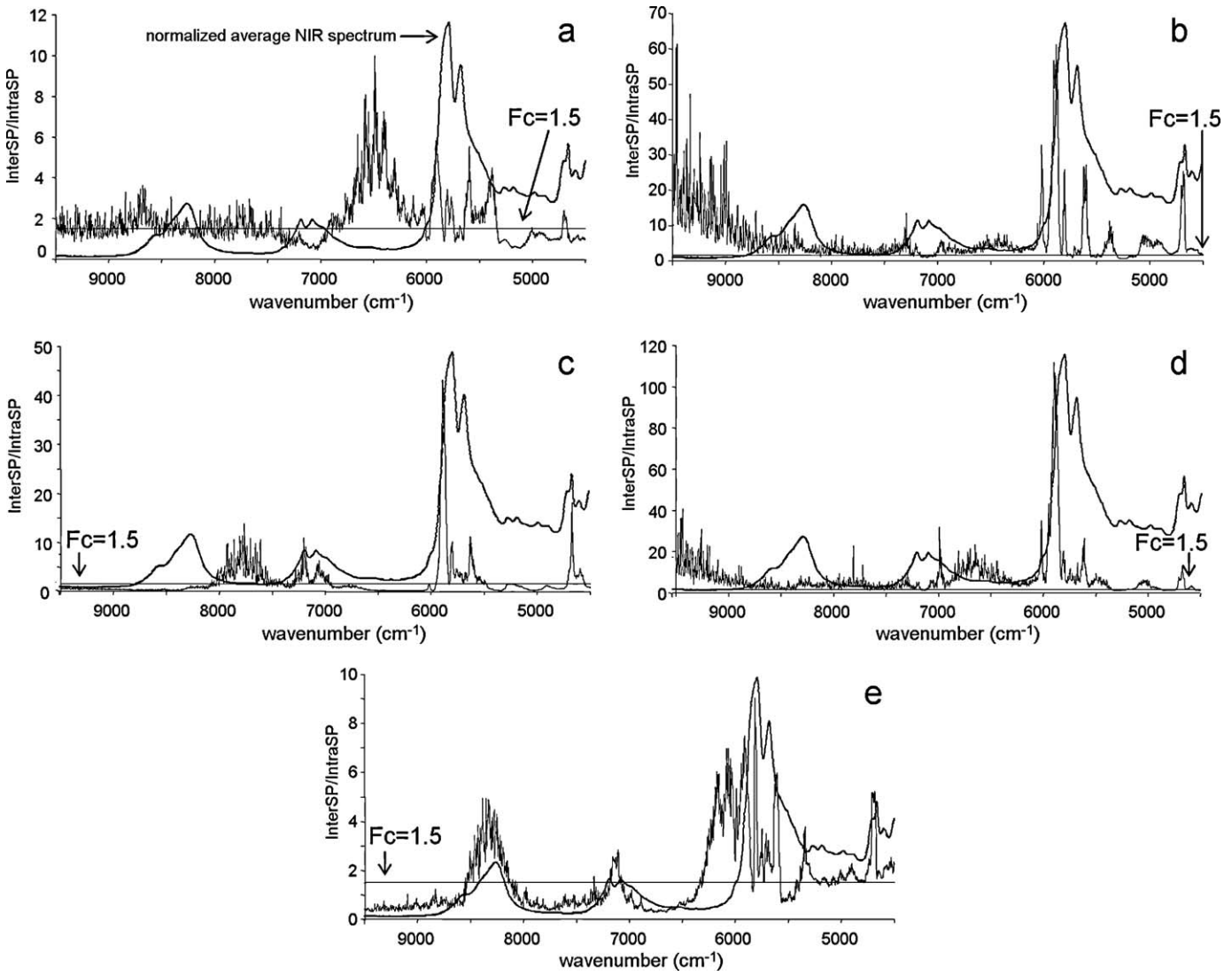

Fig. 8. Ratio of InterSP/IntraSP of each class of virgin olive oil. (a) Aix-en-Provence; (b) Haute-Provence; (c) Nice; (d) Nyons; (e) Vallée des Baux, and superposition with the mean spectra of each class and F_c.

Table 2 shows the best classification results obtained by each method to predict virgin olive oil groups according to the RDO origin (AP, HP, VB, NI and NY). Best result was obtained with PLS1-DA method, significantly better than SIMCA classification and PLS2-DA method. The results obtained using SIMCA classifications were poor for two origins (AP and VB); the percentage of correct classifications (%CC) was never more than 58% for VB and 68% for AP. The results obtained with PLS2-SIMCA are in the same order than the ones obtained with PLS1-DA, for VB and HP only one spectrum lead to bad classification.

Fig. 6a shows the %CC as a function of the number of PCs used in SIMCA model for each origin. Four components were enough to obtain the best %CC for VB prediction, but 7 PCs were used for the prediction of AP, NY and HP and 8 PCs were used for the prediction of NI. The high numbers of PCs necessary to predict origin can be explained by the overlapping showed in the PCA score plots.

Fig. 6b shows the %CC as a function of the number of LVs used in the PLS2-DA model. The best model contained at least 8 LVs. More LVs did not significantly increase the %CC.

Fig. 6c shows the %CC as a function of the number of LVs used in the PLS1-DA model. The number of LVs necessary to obtain the best model for PLS1-DA method was between 5 and 9 according to the RDO.

For all the models, prediction of the AP and VB origins were the most difficult, because these oils came from the same cultivars with

Table 2
Percentage of well classified virgin olive oils in the prediction set for each statistical treatment in the 4500–10,000 cm^{-1} spectral range.

Origins	SIMCA		PLS2-DA		PLS2-DA-SIMCA		PLS1-DA	
	%CC	PC	%CC	LV	%CC	LV-PC	%CC	LV
Aix-en-Provence (AP)	68	7	93	8	97	8–1	100	9
Haute-Provence (HP)	91	7	100	8	99	8–2	100	6
Nice (NI)	89	8	96	8	100	8–1	100	6
Nyons (NY)	89	7	100	8	100	8–1	100	2
Vallée des Baux (VB)	58	4	96	8	99	8–3	100	8

%CC: correct classification percentage; PC: principal component; LV: latent variable.

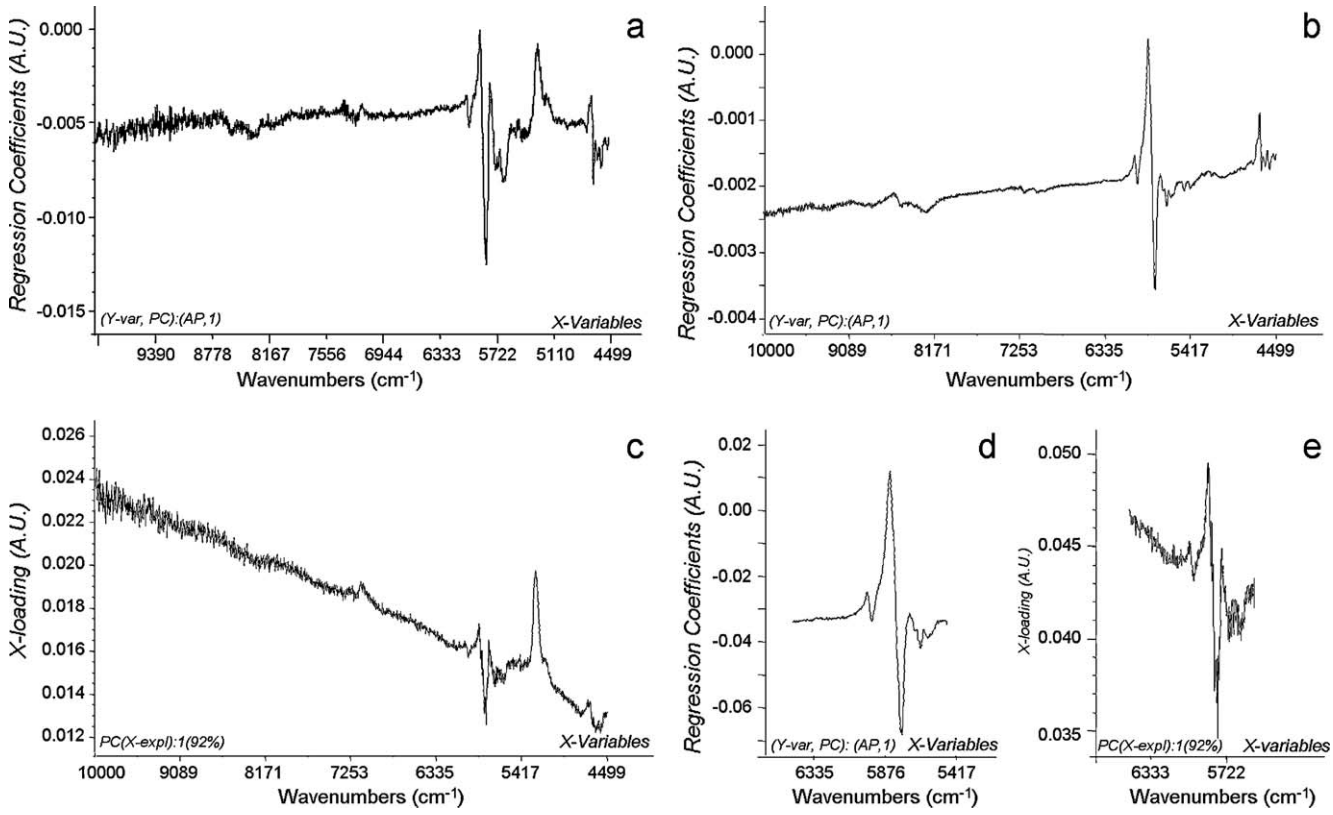

Fig. 9. First regression coefficients obtained in the 10,000–4500 cm^{-1} spectral range for AP models by (a) PLS1-DA and (b) PLS2-DA. (c) First principal component obtained for the AP group by SIMCA analysis in the 10,000–4500 cm^{-1} spectral range, (d) first regression coefficient obtained by PLS2-DA in the 6500–5500 cm^{-1} spectral range and (e) first principal component obtained for AP model by SIMCA analysis in the 6500–5500 cm^{-1} spectral range.

different rates. All the methods based on PLS-DA analysis give good results and the use of SIMCA classification performed on the scores obtained with PLS2-DA method do not improve the classification results significantly.

3.3. Classification optimization by spectral selection

According to literature [34], the variable selection conduces to improved results for quantitative analysis. The variable selection could be done on the basis of the study of spectral variance, which may directly influence classification. Two points are important, the first one is spectral variance in one group (IntraSP), the second one is spectral variance for all data (InterSP). The Fisher–Snedecor test is used to compare the two variances. Fig. 7 shows ratios of InterSP/IntraSP as function of the wavenumbers calculated for each class of crude petroleum oils, in this case according to the different degrees of freedom $F_c = 1.8$. Ratios were higher to the F_c threshold for three origins ALG, S.A., VEN (Fig. 7a, b and d), which explain good results obtained for the classification of crude petroleum oils. The spectral ranges where interSP/intraSP ratios are higher than F_c correspond to the regions with significant bands, except for the water absorption bands at the spectral ranges (3300–3100 cm^{-1}, 1700–1600 cm^{-1}, 750–700 cm^{-1}). Some crude petroleum oil samples contained a small amount of water which explained the high variance in these spectral ranges. The zones where the F_c is close to noise do not present absorption bands. For EQU (Fig. 7c), ratios were higher than F_c only in the 900–1500 cm^{-1} and 2700–3000 cm^{-1} spectral ranges. This observation is in good agreement with Fig. 2 where PCA shows high dispersion. The spectra used in this study are recorded in a spectrometer without any purge system, so it is not surprising to see some bands due to water vapor and carbon dioxide in the ratio.

Table 3
Percentage of well classified virgin olive oils in the prediction set for each statistical treatment in the 5500–6500 cm^{-1} spectral range.

Origins	SIMCA		PLS2-DA		PLS2-DA-SIMCA	
	%CC	PC	%CC	LV	%CC	LV-PC
Aix-en-Provence (AP)	70	5	100	8	96	
Haute-Provence (HP)	99	5	100	8	100	8–1
Nice (NI)	92	2	100	8	100	8–1
Nyons (NY)	95	5	100	8	100	8–1
Vallée des Baux (VB)	68	2	96	8	96	8–1

%CC: correct classification percentage; PC: principal component; LV: latent variable.

Fig. 8 shows ratios of InterSP/IntraSP as a function of wavenumbers, calculated for each class of virgin olive oils in this case according to the different degrees of freedom $F_c = 1.5$. The ratios were not always higher than the F_c threshold, which is why the SIMCA classification gives poor results. In order to confirm these results, the virgin olive oils have been classified in the 6500–5500 cm^{-1} spectral range, where InterSP versus IntraSP is higher than F_c for all virgin olive oil classes. The selected region must be the same for all the origins in order to perform PLS2 analysis. So the 6500–5500 cm^{-1} (518 points) spectral range is used for all the analyses.

Table 3 shows the best classification results obtained by each method except for PLS1-DA (100% of good predictions in the previous results) to predict virgin olive oil groups according to the RDO origin (AP, HP, VB, NI and NY). Hence, the percentage of correct classified samples increased. In the 6500–5500 cm^{-1} spectral range, SIMCA analysis performance was increased. The %CC varied between 68% and 99%. There is a clear improvement with regards to the number of PCs used in each model. The decrease of PCs used proves that models are now more robust. Both PLS2-DA and PLS2-

DA-SIMCA analysis performed a little better even if the results in Table 2 are good enough.

For AP models, Fig. 9 shows the first regression coefficient obtained by PLS1-DA (Fig. 9a), PLS2-DA (Fig. 9b) and the first principal component obtained by SIMCA analysis (Fig. 9c). It shows some variations on the spectral information used in each case. As a matter of fact, the first regression coefficient obtained by PLS1-DA (Fig. 9a) is similar to the one obtained by PLS2-DA (Fig. 9b) in the 6500–5500 cm^{-1} spectral range but it is very different in the 5500–4500 cm^{-1} spectral range. The principal component (Fig. 9c) presents high intensity in the 10,000–7000 cm^{-1} spectral range, which is correlated to instrumental deviation (baseline) and not really to chemical information. After the variable selection, the first regression coefficient obtained in PLS2-DA model (Fig. 9d) and the principal component obtained in the AP group for SIMCA analysis (Fig. 9e) are very close to the first regression coefficient obtained in PLS1-DA model (Fig. 9a).

The main difference between SIMCA and PLS-DA is the criterion used to build models. While PCA submodels are computed in SIMCA with the goal of capturing variations within each class, PLS-DA identifies directions in the data space that discriminate classes directly. Therefore, for these applications, SIMCA classification always provides worse results than methods based on PLS analysis. The difference between PLS1-DA and PLS2-DA results is significant. PLS2-DA is a version of the PLS-DA method in which several *Y*-variables are modeled simultaneously. Thus, PLS2-DA takes advantage of possible correlations or collinearties between *Y*-variables. In both cases, virgin olive oils and crude petroleum oils, the classes are independent, so the *Y* variables are less pertinent than in a PLS1-DA analysis where one model is calculated independently from the others.

4. Conclusion

In these two different applications – classification of crude petroleum oils and virgin olive oils according to their origin – PLS-DA analysis provides better results than the SIMCA method. SIMCA performance may be improved when the spectral region is reduced to the spectral range where the ratio of InterSP/IntraSP at each wavenumber is higher to the critical value of the Fisher–Snedecor test (F_c). In this case SIMCA is performed with a reduced number of principal components. PLS-DA methods are more efficient and PLS1-DA always gives better results than PLS2-DA. Therefore, the ratio InterSP/IntraSP compared to F_c is a good criterion for choosing the best infrared spectral range and the most suitable classification method for spectral analysis.

Acknowledgements

The authors are grateful to Dr Denis Ollivier (SCL) and Christian Pinatel (AFIDOL) for supplying the samples of olive oils and to Pr Albert Permanyer (University of Barcelona) for supplying the samples of crude petroleum oils.

References

[1] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.
[2] G. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.
[3] S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjostrom, H. Martens, J. Russwurm, Food Research and Data Analysis, Applied Science, Barking, 1983, p. 147.
[4] S. Wold, Pattern Recognition 8 (1976) 127.
[5] S. Wold, M. Sjöström, B.R. Kowalski, Chemometrics, Theory and Application, American Chemical Society, Washington, 1977, p. 243.
[6] O. Galtier, N. Dupuy, Y. Le Dréau, D. Ollivier, C. Pinatel, J. Kister, J. Artaud, Anal. Chim. Acta 595 (2007) 136.
[7] D. Ollivier, J. Artaud, C. Pinatel, J.P. Durbec, M. Guérère, Food Chem. 97 (2006) 382.
[8] O. Abbas, N. Dupuy, C. Rébufa, L. Vrielynck, J. Kister, A. Permanyer, Appl. Spectrosc. 60 (2006) 304.
[9] S. Perez-Magarino, M. Ortega-Heras, M.L. Gonzalez-San Jose, Z. Boger, Talanta 62 (2004) 983.
[10] J.C. Tewari, J.M.K. Irudayaraj, J. Agric. Food Chem. 53 (2005) 6955.
[11] M.P. Derde, D.L. Massart, Chemometr. Intell. Lab. Syst. 4 (1988) 65.
[12] F. Estienne, L. Pasti, V. Centner, B. Walczak, F. Despagne, D.J. Rimbaud, O.E. Noord, D.L. Massart, Chemometr. Intell. Lab. Syst. 58 (2001) 195.
[13] I.E. Frank, J.H. Friedman, Technometrics 35 (1993) 109.
[14] I.E. Frank, S. Lanteri, Chemometr. Intell. Lab. Syst. 5 (1989) 247.
[15] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Introduction to Multi- and MegaVariate Data Analysis using Projection Methods (PCA and PLS), Metrics AB, Umeå, Sweden, 1999.
[16] K.H. Esbensen, MultiVariate Data Analysis—In Practise, Camo ASA, Oslo, Norway, 2000.
[17] C. Bauer, B. Amram, M. Agnely, D. Charmot, J. Sawatzki, N. Dupuy, J.P. Huvenne, Appl. Spectrosc. 54 (2000) 528.
[18] M. Fuller, P.R. Griffiths, Anal. Chem. 50 (1988) 1906.
[19] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193.
[20] H. Martens, Anal. Chim. Acta 112 (1979) 423.
[21] Y.L. Liang, O.M. Kvalheim, Chemometr. Intell. Lab. Syst. 32 (1996) 1.
[22] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[23] A. Pedro, M. Ferreira, Anal. Chim. Acta 595 (2007) 221.
[24] S. Roussel, V. Bellon-Maurel, J.M. Roger, P. Grenier, J. Food Eng. 60 (2003) 407.
[25] M. Sjöström, S. Wold, B. Söderström, PLS discriminant plots, in: E.S. Gelsema, L.N. Kanal (Eds.), Pattern Recognition in Practice II, Elsevier, Amsterdam, 1986, p. 486.
[26] L. Stahle, S. Wold, J. Chemometr. 1 (1987) 185.
[27] R. Vong, P. Geladi, S. Wold, K. Esbensen, J. Chemometr. 2 (1988) 281.
[28] E.K. Kemsley, Chemometr. Intell. Lab. Syst. 33 (1996) 47.
[29] P. Ciosek, Z. Brzozka, W. Wroblewski, E. Martinelli, C. Di Natale, A. D'Amico, Talanta 67 (2005) 590.
[30] D. Massart, B. Vandeginste, S. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, New York, 1988.
[31] O. Abbas, Vieillissement simulé ou naturel de la matière organique. Apport du traitement chimiométrique des données spectroscopiques. Conséquences environnementales et aide à l'exploitation pétrolière, Ph.D Thesis, University P. Cézanne Aix Marseille III, 2007.
[32] P. Hourant, V. Baeten, M.T. Morales, M. Meurens, R. Aparicio, Appl. Spectrosc. 54 (2000) 1168.
[33] A. Riaublanc, D. Bertrand, E. Dufour, Lipides. In: La spectroscopie infrarouge et ses applications analytiques. Tech & Doc, Paris, 2006, p. 141.
[34] J.A.F. Pierna, O. Abbas, V. Baeten, P. Dardenne, Anal. Chim. Acta 642 (2009) 89.