Optical path length and wavelength selection using Vis/NIR

spectroscopy for olive oil's free acidity determination

Juan Francisco García Martín*

Instituto de la Grasa, Spanish National Research Council, Campus Universidad Pablo de Olavide, 41013 Seville, Spain.

Summary

Four different optical path lengths, namely 0.5, 1, 5 and 10 mm, were assayed for olive oil's free acidity determination using Vis/NIR spectroscopy. Results illustrated that the use of higher path length during spectra acquisition resulted in more accurate PLS models, especially when using solely the NIR region. The PLS model obtained with the NIR spectrum using the 10 mm cuvette was subjected to optimization by Monte Carlo uninformative variable elimination (MCUVE) and successive projections algorithm (SPA). Both methods drastically reduced the number of spectral variables and markedly improved the performance of the PLS model, especially the SPA-PLS model, which achieved a SEP (0.051) quite close to SEL (0.048). Interestingly, only 12 of the 85 spectral variables selected by SPA were among the 314 variables provided by MCUVE. All in all, NIRS incorporated to MCUVE-PLS or SPA-PLS may be applied as an alternative method for the rapid determination of olive oil's free acidity.

keywords acidity, NIR spectroscopy, olive oil, optical path length, wavelength selection.

Corresponding author. Tel.: +34 954616790; E-mail address: jf.garcia.martin@csic.es

1 **Introduction**

2

3 Olive oil is produced from olive fruits, *Olea europaea* L., in olive mills. The highest

4 quality olive oils are extracted from olive fruits solely by using mechanical or other

5 physical methods. Among the different quality parameters, free acidity is a key one for

6 olive oil classification into different levels of commercial quality. The acidity value (%)

7 accounts for free fatty acids according to its hydrolytic deterioration, so that olive oils

8 with acidity higher than 2% cannot be directly consumed and must be refined. Briefly,

9 according to the European Regulation (EUC, 1991) the maximum level of free acidity

10 for extra-virgin olive oils and virgin olive oils are 0.8% and 2%, respectively. The

11 maximum free acidity of olive oils (blends of refined olive oil and virgin olive oil fit for

12 consumption) and olive-pomace oils (obtained by treating olive pomace with solvents)

13 is 1%. The conventional determination of free acidity (EUC, 1991; 1997) must be

14 carried out in a laboratory using chemicals, being relatively costly and time-consuming.

15 An alternative non-destructive and rapid method for acidity determination could lead to

16 an early classification of olive oils in the olive mills along with an important cost

17 reduction in chemical analysis (Cayuela *et al.*, 2009).

18 The requirements of reliability, accuracy, consistency and simplicity for

19 characterization of food products, including olive oil, have encouraged the development

20 of non-destructive techniques for online measurement of their main quality parameters

21 and properties. Among these techniques, near-infrared spectroscopy (NIRS) is widely

22 used in food and pharmaceutical industries. Its widespread use is due to the fact that this

23 technique is low-cost, safe and non-destructive, as well as suitable for online work

24 (Jiménez Marquez *et al.*, 2005). Other advantages are the minimal sample preparation

(may be performed in situ in many instances) and the relatively small amounts of sample required for analysis.

With regard to the application of NIRS in the olive oil industry, the real-time determination of oil content, olive oil stability index and humidity have been previously assayed (García *et al.*, 2005; Cayuela *et al.*, 2013). Furthermore, NIR could be applied to design a quality control system for identification of olive oils adulterated with other oils such as corn oil, sunflower oil, and raw olive residue (Wesley *et al.*, 1995) and to the differentiation of olive oils which belong to different denominations of origin (Bertran *et al.*, 2000). Regarding free acidity of olive oil, its NIR prediction has been previously assayed (Jiménez Marquez *et al.*, 2005; Cayuela *et al.*, 2009), reaching significantly good calibration parameters. However, the predictive errors reported seem to be a bit high in discriminating edible olive oils (free acidity < 2%). As an example, the acidity average values for 'Arbequina' and 'Picual' olive oils were 0.49 ± 0.01 and 0.33 ± 0.00 , respectively, by means of the conventional method (EUC, 1991; 1997) while NIR predictions provided mean values of 0.54 ± 0.15 and 0.37 ± 0.16 (Jiménez Marquez *et al.*, 2005). Therefore, further optimization of NIRS for rapid determination of free acidity of edible olive oils would be welcome.

Partial least squares (PLS) is a powerful tool for NIR full-spectrum calibration. However, PLS models built using the full spectrum may contain many spectral variables with irrelevant information like noise and background (Bassbasi *et al.*, 2014; Han *et al.*, 2008). In spite of the fact that these spectral variables can apparently improve the calibration's statistics, they can worsen the predictive ability of the model. The difficulty in NIR analysis is brought about by the broad array of chemical compounds present in the sample (Shenk *et al.*, 2001), which results in an extensive overlapping of

absorption bands. Among the proposed methods for wavelength selection in multivariate calibration, Monte Carlo uninformative variable elimination (MCUVE) and the successive projections algorithm (SPA) are promising methods in chemometrics. The variable selection in MCUVE is performed according to the stability of each spectral variable in NIR analysis (Cui *et al.*, 2008; Han *et al.*, 2008; Li *et al.*, 2014). The stability is calculated with the PLS regression coefficients obtained from a large number of sample subsets generated by the Monte Carlo technique. By contrast, SPA builds subsets of variables with small collinearity using the calibration data set according to a sequence of projection operations of the columns of the instrumental response matrix (Araújo *et al.*, 2001). These variable subsets are then evaluated based on the prediction performance of the resulting multiple linear regression (MLR) models (Araújo *et al.*, 2001; Hao *et al.*, 2011).

On the other hand, there is a lack of papers concerning the effect of optical path length in NIRS. Optical path length has a significant influence on light absorption intensity at each wavelength. NIR radiation penetrates deeper into organic samples than UV, visual or far and mid infrared radiations do (García-González *et al.*, 2013), thus allowing spectral data acquisition from thick samples (i.e., 1-5 mm). Therefore, assessment of the optical path length could improve the correlation of spectral variables with analytical parameters.

The aim of this work was to determine the optimal optical path length and spectral wavelengths to build a reliable PLS model for free acidity prediction in olive oils. For these purposes, quartz cuvettes of path lengths 0.5, 2, 5 and 10 mm were tested. Calibrations were constructed with the 800-2200 nm NIR spectrum and with combinations Vis/NIR in the range of 350-2200 nm. Then, MCUVE and SPA variable

selection methods were assayed with the data obtained with the cuvette and the spectrum that led to the best model performance. The final goal was to study the feasibility of using Vis/NIR spectroscopy to determine olive oil's free acidity.

Materials and methods

Olive oil samples

A group of 127 samples from the variety 'Picual' was extracted in the laboratory of the Instituto de la Grasa by means of a MC2 mill (Ingeniería y Sistemas, S.L., Spain) from samples provided by a research project.

Free acidity reference analysis

Free acidity (FA) was expressed as oleic acid percentage and analysed according to the Official Methods of Analysis of the EC (EUC, 1991; 1997). Briefly, 4 to 6 g olive oil were placed into 250-ml wide-mouth Erlenmeyer flasks along with 50 ml ethyl alcohol:ethyl ether solution (1:1 v/v) and a few drops of phenolphthalein, and then neutralized with NaOH 0.1 N until pink in colour. The FA contents ranged between 0.10% and 1.25%. The standard error of laboratory (SEL) was 0.048%.

Instrumentation

A Vis/NIR Labspec Pro model LSP 350-2500P (Analytical Spectral Devices Inc., Boulder, CO) spectrophotometer equipped with three detectors was used for spectral

acquisition. The detector for the range 350–1000 nm is a fixed reflective holographic diode array with 512 pixels sensitivity. The wavelength range 1000–1800 nm is covered by a holographic fast scanner InGaAs detector cooled at -25°C . The same device coupled with a high order blocking filter operates for the 1801–2500 nm interval. The instrument is equipped with internal shutters and automatic offset correction, the scanning time being 100 ms. The spectrometer is equipped with a spectrophotometric cuvette accessory joined by fiber optic connectors to the light source spectrometer at one side of the accessory, and to the detector on the opposite side.

Spectral measurements

The temperature of a body has an important influence on the NIR radiation it reflects and absorbs; hence it constitutes a decisive factor in NIRS. Therefore, prior to spectra acquisition, the sample containers were placed in a thermostatic water bath fixed at 32°C for 30 min, waiting to check the stability of temperature. Spectra acquisition was carried out in transmittance mode using 4 quartz cuvettes of different path length, namely 0.5, 2 mm, 5 mm and 10 mm, respectively, with wavelength increment of 1 mm in the wavelength region from 350 to 2200 nm (1851 spectral variables). The olive oil spectra obtained were recorded using the Indico Pro software (Analytical Spectral Devices Inc., Boulder). Two replicas of each sample were acquired.

Calibration procedure

Before calibration, the reflectance data was maximum normalized and then transformed to absorbance. The resulting spectra obtained with each cuvette were

divided into calibration and validation sets by the Kennard-Stone method (Kennard & Stone, 1969; Li *et al.*, 2014) using Matlab 7.11.0.564 (R2010b) software (The MathWorks, Inc., USA). 87 samples were used for multivariable calibration and the 40 left samples were used as validation set. Partial least squares (PLS) models using full-cross internal validation (FCV) were built with The Unscrambler (CAMO Software AS, Norway). The Vis/NIR region from 350 to 2200 nm was screened with a step of 50 nm for full spectrum PLS model building.

Monte Carlo combined with uninformative variable elimination (MCUVE)

MCUVE was used for variable selection as a combination of the uninformative variable elimination (UVE) and Monte Carlo (MC) methods. Wavelengths are directly selected in MCUVE by their stability instead of adding random noise variables to the original data matrix as in UVE method for cutoff threshold estimation. The stability values are obtained through the Monte Carlo method. Therefore, the risk of over-fitting decreases and, accordingly, the evaluation of the reliability of each variable is more reasonable.

MCUVE was assayed for variable selection in the spectra obtained with the 10-mm path-length cuvette in 800-2200 nm NIR region (1401 spectral variables) and was carried out with Matlab 7.11.0.564 (R2010b) software (The MathWorks, Inc.). Following the procedure proposed by Li *et al.* (2014), the calibration set (87 samples) was randomly divided by the Monte Carlo technique into 1000 calibration subsets of 65 samples each one for constructing 1000 PLS sub-models. Then, a matrix of the PLS regression coefficients β (1000×1401) was calculated and the stability s_j of each variable j was defined as follows:

$$s_j = \frac{\text{mean}(\beta_j)}{\text{SD}(\beta_j)} \quad j = 1 \dots 1401$$

where $\text{mean}(\beta_j)$ and $\text{SD}(\beta_j)$ stand for the mean and the standard deviation of the regression coefficients of the variable j .

Successive projections algorithm (SPA)

The SPA method was also applied to the data obtained with the 10-mm path-length cuvette in the 800-2200 nm NIR region and was performed with Matlab 7.11.0.564 (R2010b) software (The MathWorks, Inc.). SPA is a forward variable selection method described elsewhere (Araújo *et al.*, 2001). SPA performs simple projection operations in a vector space to select wavelengths whose information is minimally redundant. Briefly, a first step consisted of projections carried out on the X calibration matrix ($X = N \times J$; $N = 87$, $J = 1401$), which generated J chains of variables. Each element in a chain was selected in order to display the least collinearity with the previous ones (Araújo *et al.*, 2001). The second step was choosing the optimal variable subset from the chains generated in the first phase, which was determined on the basis of the smallest root mean square error of cross validation (RMSECV) in calibration set by building multiple linear regression (MLR) models using the selected wavelengths. The subset of optimal variables was then used to build a PLS model.

Model evaluation

The performance of the models were evaluated in terms of root mean square error of calibration (RMSEC) and multiple correlation coefficient of calibration (r_c^2), and root

mean square error of prediction (RMSEP), standard error of prediction (SEP) and ratio of performance to deviation (RPD), respectively. Generally, an ideal model should have the higher r^2_c and RPD and the lower RMSECV and RMSEP (and SEP) values. Table 1 summarises the criteria proposed by Shenk & Westerhaus (1996) to assess the statistic results from calibrations and their validations, which were used throughout this research.

Results and discussion

Features of the olive oil Vis/NIR spectrum

Olive oil spectra acquired with the four cuvettes are shown in Fig. 1. Regarding the visible spectrum, a first peak occurred at about 415 nm. This area corresponds to the wavelengths of oil absorption which are dark blue colored light, and could be due mainly to carotenoids, as well as to pheophytin a, pheophorbide a and pyropheophytin a (Moyano *et al.*, 2008). A second peak near 450 nm was found, corresponding to the absorption of the blue light, characteristic of carotenoids (Moyano *et al.*, 2008). A third absorption peak at 480 nm could not be identified. A strong last peak was observed approximately at 670 nm, which coincided with the absorption of chlorophylls (Cayuela Sánchez *et al.*, 2013). It is worth noting that the first three cited peaks (between 350 and 500 nm) were easier to distinguish with the 0.5-mm and 2-mm cuvettes (Fig. 1.a and Fig. 1.b) than with the 5-mm path-length cuvette (Fig. 1.c). They could not be clearly differentiated using the 10-mm path-length cuvette (Fig. 1.d), which could indicate that

1 increasing the path length of the cuvette is in detrimental of the quality of the visible
2 spectrum of olive oil.

3 NIR spectra showed various overlapping peaks. This is because NIR spectrum is
4 sensitive to the concentrations of organic materials and these bands are the result of
5 overtones (first and second) and a combination of fundamental, largely hydrogenic,
6 vibrations that occur in the MIR region (Cayuela Sánchez *et al.*, 2013; García-González
7 *et al.*, 2013; Hao *et al.*, 2011). According to Sato *et al.* (1991), there are two regions of
8 the NIR spectrum that are of major importance. First, an absorption intensity near 1720
9 nm that is related to the first overtone of the C-H vibration of several chemical groups
10 ($-\text{CH}_3$, $-\text{CH}_2$). Second, the absorption band of the C-H vibration of *cis*-unsaturation in
11 the area of 2143 nm. Both absorption areas can be clearly identified in Fig. 1. The first
12 was found at 1727 nm. This fact is logical since the main triglyceride and the major
13 component of olive oil is triolein, and the maximum point of this band in the triolein
14 spectrum is observed at 1725 nm (García-González *et al.*, 2013). These authors (García-
15 González *et al.*, 2013) reported a maximum absorption band in the vicinity of 1724 nm
16 for the NIR spectra of olive oils. The second was found at 2144 nm (Fig. 1). Peaks
17 related to saturated and *trans* fatty acids that are usually found at 2128 nm and 2131 nm,
18 respectively (García-González *et al.*, 2013), were hardly observed. Finally, a broad
19 absorbance band occurred at around 1210 nm, probably due to second overtones of C-H
20 and CH=CH- stretching vibrations (Cayuela Sánchez *et al.*, 2013). Contrary to that
21 occurred in the visible spectrum, higher cuvette path lengths led to higher absorbances
22 in the NIR region and thereby the 10-mm quartz cuvette provided smoother NIR
23 spectra.

Vis/NIR PLS models using 0.5-mm path-length cuvette

Calibration statistics for NIR and Vis/NIR prediction models with different wavelength intervals using the 0.5-mm path-length cuvette are shown in Table 2. The different Vis/NIR models did not show any pattern and quite similar results were achieved. The r^2_c values were low, ranging from 0.42 to 0.47. They were even lower when the full Vis/NIR spectrum was assayed (0.30 for the 400-2200 nm spectrum; 0.21 for the 350-2200 nm spectrum). As for prediction, results were not satisfactory. SEP values ranged between 0.15 and 0.18, while RPD values were generally lower than 1 (maximum value of 1.01 for the 750-2200 nm spectrum). The poor results obtained could be due to the low absorbances registered, mainly in the NIR region, using the 0.5-mm quartz length cuvette (Fig. 1.a).

Vis/NIR PLS models using 2-mm path-length cuvette

NIR and Vis/NIR models for the spectra obtained with the 2 mm cuvette are shown in Table 3. Similarly to that occurred with the 0.5 mm cuvette, the addition of the visible spectrum with a step of 50 nm to the NIR spectrum did not improve the models to a great extent. Maximum RPD and r^2_c were achieved with the 400-2200 nm Vis/NIR spectrum (1.81 and 0.80, respectively), which could be considered low. RMSEC, RMSEP and SEP values ranged between 0.095 and 0.110. As SEL was 0.048, and taking into account the r^2_c values, the models obtained with the 2-mm path-length quartz cuvette could be regarded as acceptable, according to the criteria shown in Table 2. The main drawback of using the 2 mm cuvette could be that fact that the absorbance in NIR was much lower than in the visible spectrum (Fig 1.b). Therefore, the almost nil

absorbance differences among NIR spectra could be a hindrance to construct a reliable PLS calibration with this cuvette.

Vis/NIR PLS models using 5-mm path-length cuvette

The performance of the PLS models obtained with the 5 mm cuvette was clearly much higher than that of the 2 mm cuvette (Table 4). The abnormal values for the 600-2200 nm and 650-2200 nm spectra are due to the optimal PC selection made by The Unscrambler software, based on the variance explanation. When 15 PC were selected instead of 10 PC for these spectra, r^2_c , RMSEC and RMSEP were 0.952, 0.047 and 0.999 for the 600-2200 nm spectrum and 0.956, 0.045 and 0.104 for the 650-2200 nm spectrum, which are in agreement with data of Table 4. It can be observed that better calibrations were achieved when increasing the number of PC for the PLS models (PC = 15). It is worth to note that r^2_c does not indicate whether a regression model is adequate because r^2 cannot determine whether the predictions are biased. In our case, better predictions (lower RMSEP and SEP and, therefore, higher RPD) were achieved with 10 PC although the calibrations were worse. Typically, a model is overfitting the data when the quality of the prediction decreases when the number of latent variables increases, i.e. the information that fits the calibration set is not useful to fit new observations (validation set). Therefore, the PLS models obtained using 15 PC could be regarded as overfitted.

Vis/NIR PLS models using 10-mm path-length cuvette

The best performance was achieved with the 800-2200 nm spectrum (Table 5). The coefficients for model evaluation were $r^2_c = 0.974$, RMSEC = 0.050, RMSEP = 0.081 and SEP = 0.075. These results improved those found in literature. Cayuela *et al.* (2009) obtained r^2_c and RMSEC values of 0.89 and 0.13, respectively, using the 1100-2300 nm NIR spectrum and 5 mm optical path length. Jiménez Marquez *et al.* (2005), achieved r^2_c and SEP values of 0.998 and 0.16, respectively, working with the full 1100-2500 nm spectrum and 1 mm optical path length. Taking into account the criteria of Table 1, the PLS model obtained with the 10 mm cuvette can be regarded as, at least, a good precision one. However, it could be reasonable to think that optimising this NIR spectrum by variable selection would lead to better performance. Therefore, the following step was to eliminate uninformative spectral variables from the model.

Variable selection by MCUVE from NIR spectra using 10-mm path-length cuvette

The main target of this method is to determine the number of variables that leads to a robust and accurate method, hence the variation of the RMSEC of the calibration set and the RMSEP of the validation set for different variable groups was assessed. Fig. 2 depicts the RMSEC and RMSEP obtained for sets of from 50 to 1851 spectral variables with an initial step of 50 variables. From 250 to 350 variables the step was reduced to 10 spectral variables and from 300 to 320 the addition of each new spectral variable was individually assessed. Spectral variables were selected in descending order of reliable index, i.e. the first variable set was composed of the 50 wavelengths with the highest reliable indexes, and so on. For each set of selected variables, a PLS calibration model was developed which was used to predict the validation set. The minimum RMSEC was achieved with 313 spectral variables (0.055) whereas the best RMSEP was obtained by

1 using 314 spectral variables (0.073). Then both parameters increased to some extent
 2 (Fig. 2). However, from about 750 spectral variables the addition of new variables to the
 3 model improved both RMSEC and RMSEP. As a result, the PLS model of the full
 4 spectrum (1401 spectral variables) reached values close to those obtained with the
 5 optimal spectral variables selected by MCUVE (Table 6). That is, the addition of
 6 uninformative spectral variables improved the validation and, mainly, calibration
 7 models (Fig. 2). Bassbasi *et al.* (2014) illustrated how a calibration method can be
 8 inappropriate. The addition of a random new column of NIR wavelength data that has
 9 no useful information for describing their protein and moisture PLS models resulted in
 10 more effective models. In an ideal analysis, this random variable should be removed
 11 during the model-building phase Bassbasi *et al.* (2014). Unfortunately, even noise has a
 12 great possibility of being used to build a PLS method, and acidity seems to be strongly
 13 influenced by noise. Thus, Hao *et al.* (2011) assayed different wavelength selection
 14 methods (in the range 350 to 1800 nm) to determine total acidity of navel oranges.
 15 Using the full spectrum (1451 variables), RMSECV and RMSEP were 0.041 and 0.042,
 16 respectively. The application of the MCUVE method improved the prediction (RMSEP
 17 = 0.035) but not the calibration (RMSECV = 0.043) (Hao *et al.*, 2011). This accounts for
 18 the difficulty of performing variable selection for acidity and the great influence of
 19 uninformative spectral variables in performing calibration PLS models.

20 The best performance was obtained in the PLS model built with 314 variables
 21 selected by the MCUVE method, being $r^2_c = 0.931$, SEP = 0.064 and RPD = 3.01 (Table
 22 6). Taking into account the criteria of Table 2, the MCUVE-PLS model for olive oil's
 23 free acidity can be regarded as an excellent precision model. Finally, it was worth
 24 noting that most of the selected variables by MCUVE could be grouped in spectral
 25 ranges.

1

2 **Variable selection by SPA from NIR spectra using 10-mm path-length cuvette**

3 The optimal number of spectral variables, which was determined on the basis of the
4 smallest RMSECV from MLR models built in the calibration set with the wavelengths
5 proposed by the SPA method, was 85 (RMSECV = 0.034). The 85 selected variables
6 are marked in Fig. 3. Contrary to MCUVE, which mainly provided spectral ranges, SPA
7 selected individual wavelengths. Besides, the number of selected variables by SPA (85)
8 was much lower than that obtained with MCUVE (314). This is in agreement with the
9 results obtained by Hao *et al.* (2011). In their research, SPA and MCUVE selected 14
10 and 710 spectral variables, respectively, for the determination of total acidity of navel
11 oranges. These authors (Hao *et al.*, 2011) pointed out that individual wavelengths,
12 instead of spectral ranges, are more suitable to develop more parsimonious models and
13 portable instruments. However, the performance of their SPA-PLS model was slightly
14 worse than that of the full-spectrum PLS model. By contrast, the SPA-PLS model for
15 FA determination in olive oils improved the performance of the full-spectrum PLS
16 model (Table 6) and SEP (0.051) was quite close to SEL (0.048), which accounts for
17 the excellent precision of the obtained SPA-PLS model (RPD = 3.80).

18 Finally, of note is that only 12 of the 85 spectral variables selected by SPA were
19 among the 314 variables provided by MCUVE. Most of the variables selected by
20 MCUVE were related to the structure of fatty acids (aliphatic chains), which are
21 responsible for the free acidity of olive oil. However, these bands assigned to aliphatic
22 chains could also belong to triacylglycerols (the major components of olive oil).
23 Nevertheless, García-González *et al.* (2013) showed the relevant NIR wavelengths of
24 several lipids and bands that are correlated with some chemical indices (coefficient of

correlation greater than 0.90). In this way, the first overtone of triolein (cis C18:1) was found at 1725 nm and its combination band at 2142 nm, and so on. None of vibration bands assigned to lipids present in olive oils (triolein, trilinolein, tricaprin, etc.) coincided with the spectral bands provided by MCUVE. By contrast, SPA primarily focused on the wavelengths near the maximum point of absorption of triolein spectrum (1725 nm). Thus, 31 of the 85 variables selected by SPA were in the range 1711-1768 nm. Individual wavelengths such as 2141 nm and 2146 nm were selected by SPA as well, which could be correlated to the combination bands of triacylglycerols.

Conclusions

NIR spectroscopy combined with MCUVE-PLS or SPA-PLS models resulted in a powerful tool for the determination of free acidity in olive oils. Most of selected spectral region by the Monte Carlo uninformative variables elimination method could be related to the main NIR absorption bands of free fatty acids. On the contrary, SPA selected many wavelengths correlated to triacylglycerols. The increase of the optical path length in data acquisition resulted in more reliable PLS models. Results illustrated that absorption bands in the NIR spectrum had higher absorption intensity when using the 5-mm and, mainly, the 10-mm path-length quartz cuvette. On the contrary, lower optical path length seemed to work better for the visible region. While the PLS model obtained using the full NIR spectra (1401 variables) acquired with the 10-mm quartz-length cuvette provided a RPD value 2.55, the MCUVE-PLS model (314 spectral variables) increase RPD to 3.01. The SPA-PLS model further improved this performance (RPD = 3.80) and achieved a SEP value (0.051) quite close to SEL (0.048). The overall results

illustrated the potential of NIR spectroscopy as a fast and accurate method for the online determination of free acidity of edible olive oils.

Acknowledgements

This work was funded by the Recupera 2020 project (Ministry of Economy and Competitiveness, Spain).

References

- Araújo, M.C.U., Saldanha, T.C.B., Galvão, R.K.H., Yoneyama, T., Chame, H.C. & Visani, V. (2001). The Successive Projections Algorithm for variable selection in spectroscopic multicomponent. *Chemometrics and Intelligent Laboratory Systems*, **57**(2) 65–73.
- Bassbasi, M., De Luca, M., Ioele, G., Oussama, A., Ragno, G. (2014). Prediction of the geographical origin of butters by partial least square discriminant analysis (PLS-DA) applied to infrared spectroscopy (FTIR) data. *Journal of Food Composition and Analysis*, **33**, 210–215.
- Bertran, E., Blanco, M., Coello, J., Iturriaga, H., MasPOCH, S. & Monteliu, I. (2000). Near infrared spectrometry and pattern recognition as screening methods for the authentication of virgin olive oils of very close geographical origins. *Journal of Near Infrared Spectroscopy*, **8**, 45–52.

- 1 Cai, W., Li, Y. & Shao, X. (2008). A variable selection method based on uninformative
2 variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics
3 and Intelligent Laboratory Systems*, **90**, 188–194.
- 4 Cayuela, J.A., García, J.M. & Caliani, N. (2009). NIR prediction of fruit moisture, free
5 acidity and oil content in intact olives. *Grasas y Aceites*, **60**(2), 194–202.
- 6 Cayuela Sánchez, J.A., Moreda, W. & García, J.M. (2013). Rapid Determination of
7 Olive Oil Oxidative Stability and Its Major Quality Parameters Using Vis/NIR
8 Transmittance Spectroscopy. *Journal of Agricultural and Food Chemistry*, **61**(34),
9 8056–8062.
- 10 García, A., Ramos, N. & Ballesteros, E. (2005). Estudio comparativo de distintas
11 técnicas analíticas (espectroscopía de NIR y RMN y extracción mediante Soxhlet) para
12 la determinación del contenido graso y de humedad en aceitunas y orujo de Jaén.
13 *Grasas y Aceites*, **56**, 220–227.
- 14 García-González, D.L., Baeten, V., Fernández Pierna, J.A. & Tena, N. (2013). Infrared,
15 Raman, and fluorescence spectroscopies: Methodologies and applications, in Aparicio R
16 & Harwood J (Eds.) *Handbook of olive oil: Analysis and properties*. Springer, New
17 York, 335–393.
- 18 Kennard, R.W. & Stone, L.A. (1969). Computer aided design of experiments.
19 *Technometrics*, **11**, 137–148.
- 20 EUC. (1991). European Union Commission Regulation Characteristics of olive and
21 olive Pomace Oils. *Off. J. Eur. Communities*. Anexe IX.
- 22 EUC. (1997). European Union Commission Regulation nº 2472/97, amending the
23 Regulation nº 2568/91. *Off. J. Eur. Communities*.

- 1 Han, Q.-J., Wu, H.-L., Cai, C.-B., Xu, L. & Yu, R.-Q. (2008). An ensemble of Monte
2 Carlo uninformative variable elimination for wavelength selection. *Analytica Chimica*
3 *Acta*, **612**, 121–125.
- 4 Hao, Y., Sun, X., Zhang, H. & Liu, Y. (2011). Application of effective wavelength
5 selection methods to determine total acidity of navel orange. *Sensor Letters*, **9**(3),
6 1229–1234.
- 7 Jiménez Marquez, A., Molina Díaz, A. & Pascual Reguera, M.I. (2005). Using optical
8 NIR sensor for on-line virgin olive oils characterization. *Sensors and Actuators B*, **107**,
9 64–68.
- 10 Li, H.D., Xu, Q.S. & Liang, Y.Z. (2014). libPLS: An Integrated Library for Partial
11 Least Squares Regression and Discriminant Analysis. *Peer J PrePrints* **2**, e190v1.
- 12 Moyano, M.J., Meléndez, A.J., Alba, J. & Heredia, F.J. (2008). A comprehensive study
13 on the colour of virgin olive oils and its relationship with their chlorophylls and
14 carotenoids indexes (I): CIEXYZ non-uniform colour space. *Food Research*
15 *International*, **41**, 505–512.
- 16 Sato, T., Kawano, S. & Iwamoto, M. (1991). Near-infrared spectral patterns of fatty
17 acid analysis from fats and oils. *Journal of the American Oil Chemists' Society*, **68**,
18 827–883.
- 19 Shenk, J.S. & Westerhaus, M.O. (1996). Calibration the ISI way, in: Davies, A.M.C.,
20 Williams, P.C. (Eds.) *Near-infrared spectroscopy: The Future Waves*. NIR
21 Publications, Chichester, UK, 198–202.

Shenk, J.S., Workman, J.J. & Westerhaus, M.O. (2001). Application of NIR spectroscopy to agricultural products, in: Burns, D.A. & Ciurczak, E.W. (Eds.) *Handbook of near-infrared analysis*. Marcel Dekker, Inc., New York, 419–474.

Wesley, I.J., Barnes, R.J. & McGill, A. (1995). Measurement of adulteration of olive oils by near infrared spectroscopy. *Journal of the American Oil Chemists' Society*, **3**, 289–298.

Table 1

Criteria for the evaluation of the results

r_c^2		SEP	
$r_c^2 \geq 0.90$	Excellent precision	SEP = 1–1.5 SEL	Excellent precision
$r_c^2 = 0.70–0.89$	Good precision	SEP = 2–3 SEL	Good precision
$r_c^2 = 0.50–0.69$	Good separation between low, medium, and high values	SEP = 4 SEL	Medium precision
$r_c^2 = 0.30–0.49$	Correct separation between low and high values	SEP = 5 SEL	Low precision
$r_c^2 = 0.05–0.29$	It is better than no analysing		

r_c^2 : correlation coefficient of calibration; SEP: standard error of prediction; SEL:

standard error of laboratory

Table 2

**Calibration and prediction results of FA by PLS models with 0.5-mm path-length
cuvette**

Spectral region (nm)	r_c^2	RMSEC	RMSEP	SEP	RPD	PCs
800-2200	0.422	0.180	0.166	0.169	0.87	7
750-2200	0.417	0.173	0.152	0.145	1.01	7
700-2200	0.418	0.172	0.164	0.166	0.88	7
650-2200	0.446	0.168	0.164	0.166	0.89	8
600-2200	0.469	0.165	0.175	0.176	0.83	8
550-2200	0.445	0.168	0.176	0.176	0.83	8
500-2200	0.471	0.164	0.173	0.173	0.85	9
450-2200	0.445	0.168	0.169	0.169	0.86	8
400-2200	0.362	0.180	0.151	0.149	0.99	7
350-2200	0.210	0.200	0.152	0.147	1.00	5

r_c^2 : correlation coefficient of calibration; RMSEC: root mean square error of calibration; RMSEP: root mean square error of prediction; SEP: standard error of prediction; RPD: ratio of performance to deviation; PCs: number of principal components.

1
2
3

4
5
6
7
8
9
10
11
12
13
14

Table 3
**Calibration and prediction results of FA by PLS models with 2-mm path-length
cuvette**

Spectral region (nm)	r^2_c	RMSEC	RMSEP	SEP	RPD	PCs
800-2200	0.740	0.110	0.104	0.105	1.71	7
750-2200	0.706	0.110	0.107	0.108	1.67	7
700-2200	0.737	0.110	0.109	0.111	1.63	7
650-2200	0.737	0.110	0.109	0.110	1.64	8
600-2200	0.742	0.109	0.109	0.110	1.64	8
550-2200	0.749	0.108	0.107	0.109	1.66	8
500-2200	0.743	0.109	0.108	0.109	1.65	9
450-2200	0.763	0.105	0.107	0.108	1.67	10
400-2200	0.803	0.096	0.098	0.100	1.81	12
350-2200	0.792	0.098	0.110	0.109	1.65	11

r^2_c : correlation coefficient of calibration; RMSEC: root mean square error of calibration; RMSEP: root mean square error of prediction; SEP: standard error of prediction; RPD: ratio of performance to deviation; PCs: number of principal components.

Table 4

**Calibration and prediction results of FA by PLS models with 5-mm path-length
cuvette**

Spectral region (nm)	r^2_c	RMSEC	RMSEP	SEP	RPD	PCs
800-2200	0.965	0.040	0.112	0.102	1.73	15
750-2200	0.964	0.041	0.118	0.101	1.75	15
700-2200	0.961	0.042	0.102	0.092	1.91	15
650-2200	0.858	0.081	0.088	0.084	2.09	10
600-2200	0.852	0.083	0.096	0.093	1.90	10
550-2200	0.950	0.048	0.102	0.087	2.04	15
500-2200	0.927	0.058	0.096	0.085	2.07	14
450-2200	0.882	0.074	0.095	0.091	1.93	14
400-2200	0.864	0.080	0.086	0.085	2.09	14
350-2200	0.465	0.158	0.116	0.116	1.52	6

r^2_c : correlation coefficient of calibration; RMSEC: root mean square error of calibration; RMSEP: root mean square error of prediction; SEP: standard error of prediction; RPD: ratio of performance to deviation; PCs: number of principal components.

Table 5

**Calibration and prediction results of FA by PLS models with 10-mm path-length
cuvette**

Spectral region (nm)	r^2_c	RMSEC	RMSEP	SEP	RPD	PCs
800-2200	0.943	0.050	0.081	0.075	2.55	15
750-2200	0.939	0.052	0.081	0.076	2.54	12
700-2200	0.933	0.055	0.088	0.078	2.45	12
650-2200	0.949	0.048	0.091	0.087	2.20	14
600-2200	0.968	0.038	0.084	0.081	2.37	15
550-2200	0.934	0.050	0.084	0.078	2.45	12
500-2200	0.951	0.047	0.098	0.093	2.06	15
450-2200	0.943	0.050	0.096	0.100	1.93	15
400-2200	0.773	0.100	0.123	0.121	1.59	10
350-2200	0.404	0.163	0.135	0.134	1.43	5

r^2_c : correlation coefficient of calibration; RMSEC: root mean square error of calibration; RMSEP: root mean square error of prediction; SEP: standard error of prediction; RPD: ratio of performance to deviation; PCs: number of principal components.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Table 6
Comparison between results of FA obtained by different methods with the NIR spectra acquired with the 10-mm path-length cuvette

Method	n	r^2_c	RMSEC	RMSEP	SEP	RPD	PCs
PLS	1401	0.943	0.050	0.081	0.075	2.55	15
MCUVE-PLS	314	0.931	0.056	0.073	0.064	3.01	11
SPA-PLS	85	0.957	0.044	0.053	0.051	3.80	21

n: variable number; r^2_c : correlation coefficient of calibration; RMSEC: root mean square error of calibration; RMSEP: root mean square error of prediction; SEP: standard error of prediction; RPD: ratio of performance to deviation; PCs: number of principal components.

Figure 1. Olive oil Vis/NIR spectra obtained with 0.5-mm (a), 2-mm (b), 5-mm (c) and 10-mm (d) path-length quartz cuvettes.

Figure 2. Variation of RMSECV (dashed line) and RMSEP (continuous line) with the number of selected variables by MCUVE for the FA of olive oil.

1

2

3 **Figure 3.** Selected wavelengths by the SPA method.