ACCEPTED VERSION

On-line system based on hyperspectral information to estimate acidity, moisture and peroxides in olive oil samples

D.M. Martínez Gila *, P. Cano Marchal, J. Gámez García, J. Gómez Ortega

Robotics, Automation and Computer Vision Group, University of Jaén, Campus Las Lagunillas s/n, 23071 Jaén, Spain

ABSTRACT

The analysis of the quality of virgin olive oil involves the determination of a series of properties, such as chemical indexes and organoleptic characteristics. In addition, the determination of these properties in real-time could be useful in order to improve the olive oil extraction process since the process parameters could be regulated with the real-time moisture information.

In this paper, the feasibility of using a non-invasive hyperspectral device, in order to determine on-line three parameters of the olive oil (free acidity, peroxide index and moisture) is studied. In order to study the correlation between these parameters and the information obtained by the hyperspectral sensor (absorption level), three different methods were applied: genetic algorithms (GA), least absolute shrink-age and selection operator (LASSO), and successive projection algorithm (SPA). From the experimental results, reduced values in cross validation were obtained and the optimal wavelengths were pointed out.

1. Introduction

In general, the quality of a product is determined by the set of characteristics that allow it to be classified as equal, better or worse than those of similar nature. For olive oil, the parameters that determine its quality are influenced by the olive fruit (origin, variety, maturity and agronomic conditions, mainly) and the elaboration process variables (thermomixer temperature, centrifugal decanter differential velocity and others) (Cano et al., 2011).

Currently, the official analysis of the physicochemical and organoleptic quality parameters of olive oil are performed according to the methods established in the European norm CE 1989/03 (CE, 2003). The procedures are manual, work-intensive methods and not eligible for their on-line implementation in order to adjust the elaboration parameters automatically.

In recent years, there have been several advances related to non-invasive techniques for automatic food quality analysis (Baiano et al., 2012; Huang et al., 2012; Jamshidi et al., 2012; Suphamitmongkol et al., 2013) including olive oil (Ruiz Altisent et al., 2010; Sinelli et al., 2010). They have been focused in the infrared spectral range and have achieved good results. For instance, in El-Abassy et al. (2009) a Raman spectroscopy in the spectral window 945–1600 cm^{-1}, which includes carotenoid bands, was found to be a useful fingerprint region, being statistically significant for the prediction of the free fatty acids. Also, in Armenta et al. (2007), one FT-NIR (Fourier Transform Near Infrared) spectrophotometer device was employed for the determination of acidity and peroxides index in edible olive, sunflower seed and maize oils (Inarejos García et al., 2013; Mailer, 2004; Marquez, 2003). Moreover, in Jimenez et al. (2008) one hyperspectral device working in the spectral range between 1100 and 2300 nm, the AOTF-NIR (Acousto-Optic Tunable Filter Near Infrared) spectrophotometer, was used for real-time prediction of the moisture and fat content in olive pomace using two-phase olive oil processing. To our knowledge, there are no contributions dealing with the selection of the best wavelengths for the evaluation of olive oil quality parameters from images captured by an imaging spectrograph.

In this context, and using the images acquired by an hyperspectral device, the goal of this work was to select the optimal wavelengths which are better correlated with the olive oil parameters: free acidity, peroxide index and moisture. With this information, the computation time of the regression algorithms and the hardware costs for building new hyperspectral sensors could be significantly reduced. To this end, three methods were studied in order to select the optimal wavelengths: genetic algorithm (GA) (Eiben and Smith, 2003), least absolute shrinkage and selection operator (LASSO) (Zou and Hastie, 2005) and successive projection algorithm (SPA) (Arajo et al., 2001). To our knowledge, these methods were selected for our experimentation because they are widely used in the literature and they have achieved good results with

* Corresponding author.
E-mail address: dmgila@ujaen.es (D.M. Martínez Gila).

http://dx.doi.org/10.1016/j.compag.2015.06.002

spectral data inputs (Andersen and Bro, 2010; Arajo et al., 2001). However, in the last years other variable selection algorithms and variations (Monteiro and Kosugi, 2007; Latorre Carmona et al., 2012) have arisen that could be used in future work. The correlation between the selected wavelengths and the values of the parameters obtained by means of analysis performed in a laboratory were evaluated with multi-linear regression (MLR) (Aiken et al., 2003).

This article is structured as follows. Section 2 describes the olive oil samples used for the experiments, the analytical methods employed in order to reach the reference parameters, the experimental set-up built for acquiring images and the mathematical algorithms used for selecting wavelengths. Then, Section 3 shows the results and its comments. Finally, Section 4 presents our conclusions.

2. Materials and methods

The procedure followed in this paper has been the selection and preparation of the olive oil samples, the measurement of these samples by a hyperspectral sensor, their analysis in a certified laboratory, and the study of the resulting data.

2.1. Olive oil samples

A total of 133 olive oil samples of around 30 cl per sample were provided by the olive oil laboratory CM Europa S.L. (www.cmeuropa.com). This number included different types of olive oil -virgin, extra virgin, lampante- from the campaigns of 2012–2013 and 2013–2014. A database with acidity, peroxides levels, moisture and other chemical analysis applied to the samples were given too.

Of these samples, 56 of them included peroxides index values, 69 of them with moisture values and 133 samples with free-acidity analysis. Maximum, minimum and average values are shown in Table 1.

2.2. Analytical methods

The analytical methods were carried out by CM Europa. Thus, the acidity index is determined by acid-base titration of the free fatty acids with potassium hydroxide of the olive oil sample dissolved in ethanol. In turn, the peroxides index is determined dissolving the sample in a mixture of acetic acid and chloroform, later in a potassium iodide solution and titrating the freed iodine. The moisture content is determined relating the weight of the sample before and after a drying process held in a drying oven.

2.3. Experimental setup

The hyperspectral camera device used was composed of a Xeva-1.7-320 digital camera with a thermo-electrically cooled InGaAs detector head, an ImSpector N17E spectrograph and a 8 mm lens (ImSpector, 2003). The integration time was set in 1 ms. It behaves like a lineal camera capturing one line of the sample composed of 320 pixels over 256 different wavelengths. Its spectral range captured is between 900 nm and 1700 nm, with a resolution of 4 nm. The result is an image with 320 columns and 256 rows, equal to 81,920 pixels codified by 14 bits of resolution. The device captures up to 120 frames per second which can be transmitted to a PC through its CameraLink interface.

An automatic sampler was built with the hyperspectral camera device, a 100 W halogen lamp, a conveyor controlled by a LXM32M speed shifter and an infrared sensor used to detect the presence of objects. The setup is shown in Fig. 1. The images were captured using the external trigger of the camera connected to the infrared sensor. In measuring time the belt was stopped.

The software employed to control the LXM32M speed shifter was SoMove Lite V.1.4.4.0. On the other hand, Matlab 7.11.0 (R2013a) was used for the development of the software for the capture and analysis of the hyperspectral images.

2.4. Data analysis

The hyperspectral images were first preprocessed to obtain the spectra of the samples and also to reduce the noise introduced by the lighting system and other sources. Then, component selection algorithms were employed based on the following prefilters.

Table 1
Reference parameters from laboratory.

	Number of samples	Maximum	Minimum	Average ± SD
Acidity %	133	1.99	0.12	0.48 ± 0.45
Peroxides index meq. O_2	56	6.70	3.90	5.31 ± 0.83
Moisture %	69	0.43	0.07	0.15 ± 0.07

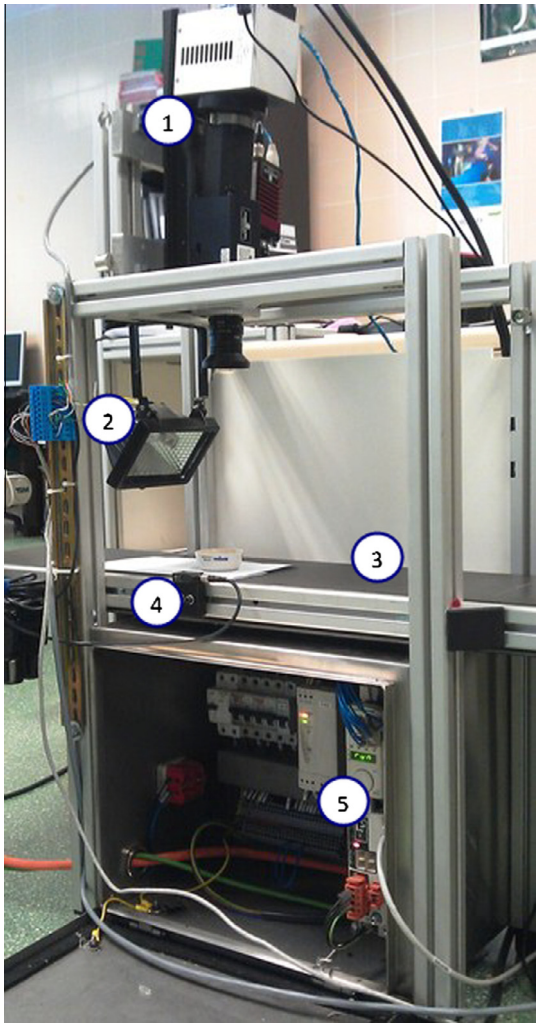


Fig. 1. Vision system configured. In the figure, label 1 indicates the InGaAs hyperspectral camera, label 2 the illumination system used, label 3 the conveyor belt, label 4 the infrared sensor and label 5 indicates the variable speed drive.

2.4.1. Hyperspectral images preprocessing

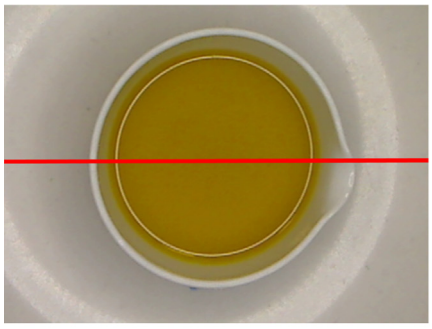
Before starting the capture of the images, the samples were prepared by pouring a small quantity of oil in a basin and placing them on the conveyor of the experimental setup over a white paper. The line of the sample which crosses the center of the basin was captured by the camera device (Fig. 2(a)). An example of captured image is shown in Fig. 2(b), where the X axis of the image belongs to a line in the sample, and the Y axis contains the wavelengths.

The spectra of the olive oil samples are coded in each column of the monochromatic image. The first step consists of eliminating the blind frequency bands that appear due to the model of hyperspectral camera employed and do not provide any information. Specifically, these are the wavelengths included between 900–994 nm and 1678–1700 nm. After this, an average spectrum is calculated according to the Eq. (1):

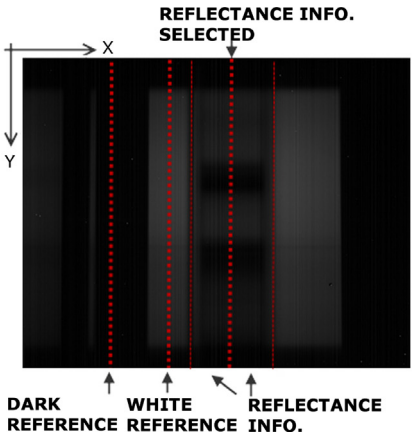
$$\overline{E}(y) = \frac{\sum_{x=x_0-\Delta x}^{x_0+\Delta x} E(x,y)}{2\Delta x} \tag{1}$$

where x_0 represents the selected pixel of the sample, Δx is half the number of pixels to average and $E(x,y)$ is the hyperspectral image. The variable x is the pixel of the line in the captured sample, and y is the hyperspectral information for the different wavelengths.

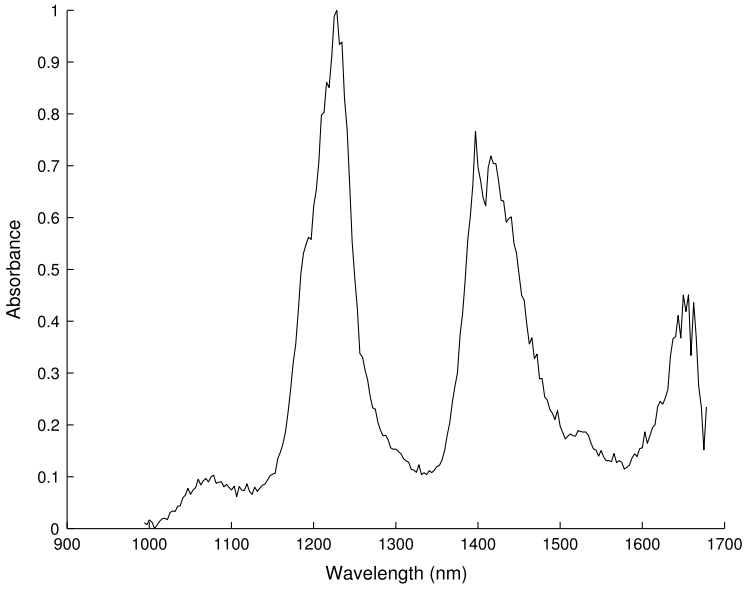
Afterwards, the image is calibrated with respect to two surfaces with non-variable spectrum. These were the white paper under the basin and the black conveyor belt. That is:



(a) Image of an olive oil sample contained in a basin. The red line denotes the piece of sample to analyze. The hyperspectral image for this line is shown in the Figure b.



(b) Image (320x256) captured by hyperspectral camera. Discontinuous lines emphasize the white reference (white paper), dark reference (conveyor belt) and olive oil reflectance information.



(c) Spectrogram without any preprocessing of an olive oil hyperspectral image.

Fig. 2. Spectra acquisition.

$$I = \frac{I_0 - D}{W - D} \times 100 \quad (2)$$

where I_0 is the spectrum to calibrate, D is the black surface spectrum and W is the white surface spectrum.

Then, the spectra were turn into energy absorbed by the oil samples according to Eq. (3).

$$A = \log \frac{1}{I} \quad (3)$$

Fig. 2(c) shows the resultant spectrum. After that, the near infrared spectra were standardized by using standard normal variate (SNV) and filtered by Savitzky–Golay (SG) method (Savitzky and Golay, 1964). The SG input parameters were tuned with the objective of achieving good correlation results. Also, the aforementioned methods were applied in order to remove and minimize any unwanted spectral contributions such as light scatter (Yao et al., 2010).

2.4.2. Wavelength selection methods

Three known component selection algorithms were evaluated with the objective of searching for the best wavelengths which are correlated with the aforementioned parameters. They were Genetic Algorithm (GA) (Eiben and Smith, 2003), Least Absolute Shrinkage and Selection Operator (LASSO) (Zou and Hastie, 2005) and Successive Projection Algorithm (SPA) (Arajo et al., 2001).

Table 2
Results for free acidity estimation.

Algorithm applied	Pre-processing method	Number of components	RMSEV[a]	R_v[b]	RPD[c]
GA-MLR	ABS, SNV	106	0.0949	0.98	4.77
SPA-MLR	ABS, SNV	106	0.1628	0.93	2.78
LASSO	ABS, SNV, SAVGOL	22	0.3114	0.84	1.45

[a] Root mean square error in validation.
[b] Linear regression coefficient in validation.
[c] Residual predictive deviation.

The genetic algorithm was coded by using the GA Toolbox of Matlab (MathWorks, 2004) in order to select the spectral components of the samples that optimize the multiple linear regression (MLR) method (Aiken et al., 2003). Specifically, for the problem treated in this work, the steady-state GA model was employed, where the population of individuals is formed by binary vectors and every component of the spectrum (220 components) is represented by one bit. If the component is selected, the value of the bit is "1", being "0" otherwise. The genetic algorithm uses a cost function in order to evaluate the selected spectral components. Individuals showing the lowest values of the cost function are more likely to be propagated to the next generation. In this case, the cost function was the mean squared error obtained with the MLR method. The parameters of the genetic algorithm empirically selected to control the convergence speed were:

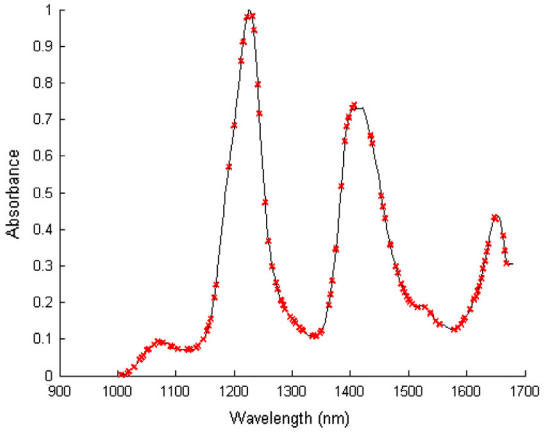
- Number of individuals per generation: 100.
- Number of generations: 70.
- The genetic algorithm stops when the cost function is not improved during 10 iterations.

On the other hand, a LASSO regression algorithm was applied to the spectra data with the goal of selecting wavelengths. LASSO uses an l_1-penalty and continuously shrinks the smallest estimated regression coefficients towards zero. The number of zero-valued regression coefficients increases as a function of the regularization parameter λ. The Eq. (4) explains the problem.

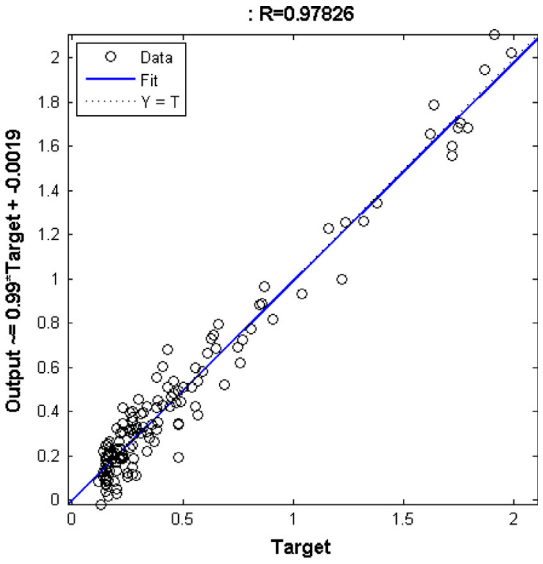
$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (4)$$

where n is the number of olive oil samples, p is the number of wavelengths (220 in this case), β_j are the regression coefficients and λ is the regularization parameter. Matlab Statistic Toolbox was used for this method (MathWorks, 2013).

Successive projections algorithm was also evaluated as selection method. This algorithm starts with one wavelength, then incorporates a new one at each iteration, until a specified number N_{max} of wavelengths is reached. Its purpose is to select a number of wavelengths whose information content is minimally redundant



(a) Components selected by GA method, marked with x marks



(b) Regression fitted by cross validation

Fig. 3. Free acidity estimation results.

and it solves collinearity problems. From $N = 1$ to N_{max} selected wavelengths, a MLR calibration model was employed in order to find the minimal mean square error of cross validation.

In summary, three models were used: GA-MLR, LASSO and SPA-MLR. The performance of calibration models was evaluated in terms of Mean Square Error of Validation (MSEV) and Residual Predictive Deviation (RPD) (Williams and Norris, 2004).

$$\mathrm{MSEV} = \frac{\sum_{i=1}^{N_v} (\hat{y}_{iv} - y_i)^2}{N_v} \tag{5}$$

being y_i the value measured in laboratory for the sample i, $\hat{y}_{iv}$ is the predicted value by validation model for the sample i, and N_v is the number of validation samples. To validate the model the leave-one-out approach was used.

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSEV}} \tag{6}$$

where SD is the standard deviation of the reference data which were measured in laboratory and RMSEV is the root of MSEV.

3. Results and discussion

Fig. 2(c) shows the raw NIR spectra of all olive oil samples. The spectra did not evidence any obvious differences from visual inspection on the basis of their compositional features. In the NIR region, bands around 1200 nm arise from 2nd overtones of C-H stretching vibrations while those at 1400 nm and 1410 nm are due to the combination band of C-H (Christy et al., 2004; Cozzolino et al., 2005).

Table 2 shows the results for acidity determination by employing different component selection algorithms. The best results were obtained with the dataset turned into absorbance levels (ABS) and standard normal variate (SNV) transformation without the Savitzky–Golay filter (SAVGOL). The last method could be unnecessary because the spectrum has been averaged by the spectral resolution of the hyperspectral device (3 nm). On the other hand, the best RMSEV value was reached with GA-MLR method. In this case, an error in the estimation of free acidity of 0.0949% was obtained and the predicted values were fitted with a regression coefficient of 0.98 (Fig. 3(b)). This result is appreciably good in order to classify on-line olive oil quality where the free acidity value between virgin and extra virgin class is in 0.8%. The components selected by GA algorithm are shown in Fig. 3(a). Most components are located between 1000–1160 nm (23% of the total), 1280–1350 nm (13%), 1480–1500 nm (5%) and 1570–1640 nm (14%). Also, the validation carried out by the orthogonal components selected by SPA-MLR method obtained good results with an error of 0.1628% (lower than the GA one). Finally, the best error obtained with LASSO method was reached with 22 components (0.3114%).

Table 3
Results for peroxide index estimation.

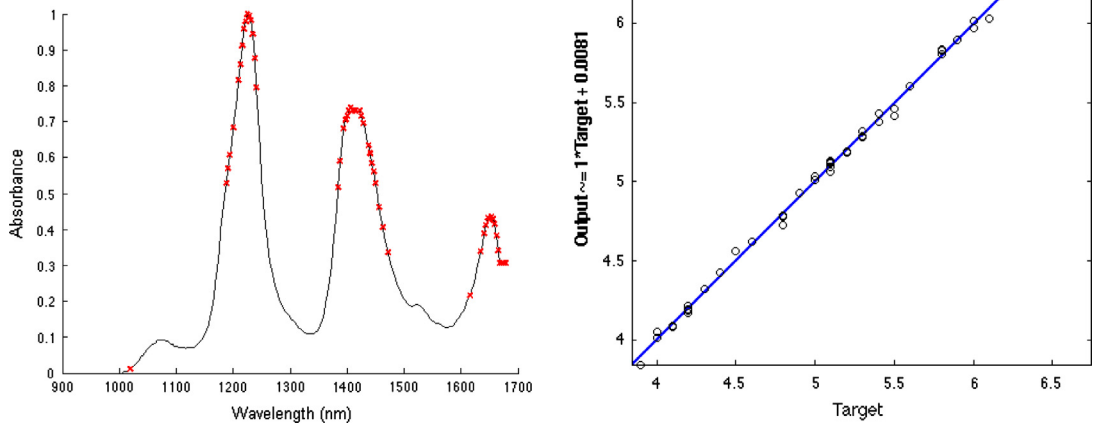
Algorithm applied	Pre-processing method	Number of components	RMSEV[a]	R_v[b]	RPD[c]
GA-MLR	ABS, SNV, SAVGOL(3,0,0)	105	0.1830	0.97	4.55
SPA-MLR	ABS, SNV	52	0.0316	0.99	26.38
LASSO	ABS, SNV, SAVGOL(7,0,0)	7	0.5280	0.84	1.58

[a] Root mean square error in validation.
[b] Linear regression coefficient in validation.
[c] Residual predictive deviation.

Table 4
Results for moisture estimation.

Algorithm applied	Pre-processing method	Number of components	RMSEV[a]	R_v[b]	RPD[c]
GA-MLR	ABS, SNV, SAVGOL(3,0,0)	103	0.0242	0.94	2.89
SPA-MLR	ABS, SNV, SAVGOL(3,3,1)	67	0.0006	0.99	116.66
LASSO		18	0.0600	0.80	1.16

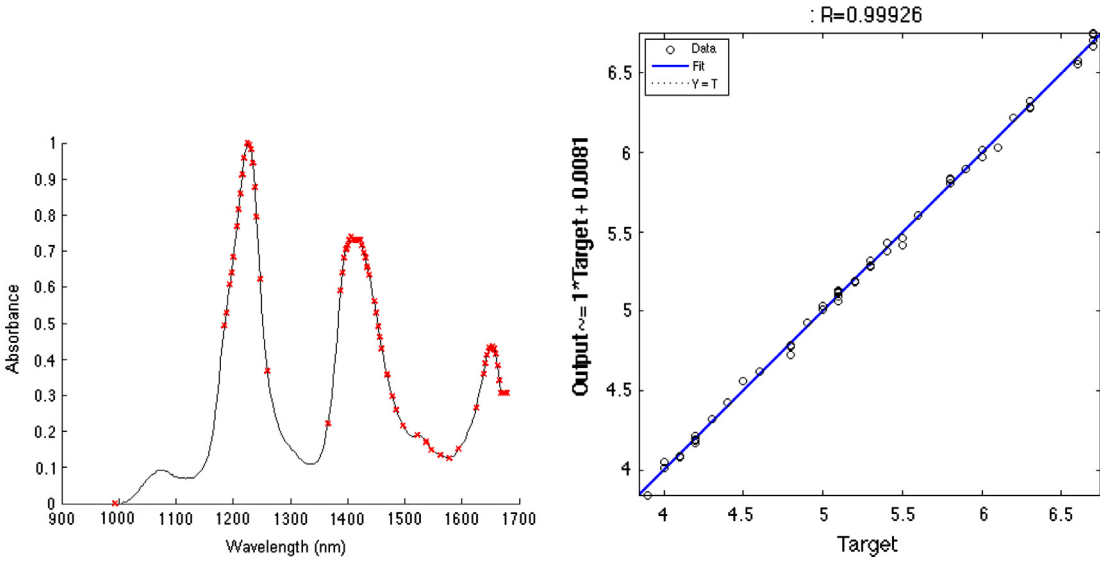
[a] Root mean square error in validation.
[b] Linear regression coefficient in validation.
[c] Residual predictive deviation.



(a) Components selected by SPA method, pointed with x marks

(b) Regression fitted by cross validation

Fig. 4. Peroxide index estimation results.



(a) Components selected by SPA method, noted with x marks

(b) Regression fitted by cross validation

Fig. 5. Moisture estimation results.

On the other hand, Table 3 presents the minimum error values obtained for each method for peroxide index determination. In this case, the best result was reached with the SPA selection method and the spectral responses transformed into ABS and SNV. The SPA with 52 components adjusted the response variables with a regression coefficient of 0.99 and 0.03 of MLR residual error (Fig. 4(b)). Fig. 4(a) shows the selected wavelength. These were grouped around the three main peaks: 11 components or (21% of the total) are located between 1210 and 1240 nm, the same quantity between 1390 and 1430 nm and 10 components (20%) between 1630 and 1670 nm.

Finally, the last table (Table 4) shows the results for moisture determination. Also, for this parameter the best result was obtained with SPA selection method and the three preprocessing methods, ABS, SNV and SAVGOL. The last one with (3,3,1) parameters, that is, mobile average with three points, 3rd degree polynomial and first derivative. These parameters reached an MLR error value of 0.0006 and a regression coefficient of 0.99 (Fig. 5(b)). The components selected by SPA algorithm are shown in Fig. 5(a). The components were distributed around the three main peaks (11 (16%) in 1206–1241 nm, 16 (24%) in 1390–1440 nm, 8 (12%) in 1640–1660 nm) and 15 wavelengths (22%) in the valley located between 1447 and 1594 nm.

4. Conclusions

This investigation has tested the feasibility of using an non-invasive hyperspectral device in order to determinate three interesting parameters in the olive oil extraction process: free acidity, peroxide index and moisture. In this study, the optimal wavelengths for the parameter estimation have been identified by using three component selection algorithms: genetic algorithm (GA), least absolute shrinkage and selection operator (LASSO) and successive projection algorithm (SPA). The best results have been achieved according to the minimum cross validation error after applying multiple linear regression (MLR). In particular, the error values obtained were 0.09% for acidity percentage, 0.03 meq $\cdot O_2$ for peroxide index and 0.0006% for moisture percentage. Considering these reduced error values and the well-fitted regression lines, the suitability of using the device in an industrial plant could be concluded. Also, the identification of the optimal wavelengths could allow the reduction of the array sensor size and thus the sensor cost. The chemical reference parameters have been taken from the analysis of the real olive oil samples carried out by the accredited laboratory of CM Europa S.L company.

Acknowledgments

This work was partially supported by the projects DPI2011-27284, TEP2009-5363 and AGR-6429. We also would like to extend a special thank to CM Europa (Martos, Spain), for their support in the lab measures and samples. Diego Manuel Martnez Gila is in receipt of a FPI grant from the Regional Government of Andalusia.

References

Aiken, L.S., West, S.G., Pitts, S.C., 2003. Multiple linear regression. In: Handbook of Psychology. John Wiley and Son.

Andersen, C.M., Bro, R., 2010. Variable selection in regression – a tutorial. J. Chemometr. 24.

Arajo, M.C.U., Saldanha, T.C.B., Galvo, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. Syst. 57, 65–73.

Armenta, S., Garrigues, S., de la Guardia, M., 2007. Determination of edible oil parameters by near infrared spectrometry. Anal. Chim. Acta 596, 330–337.

Baiano, A., Terracone, C., Peri, G., Romaniello, R., 2012. Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. Comput. Electron. Agric. 87, 142–151.

Cano Marchal, P., Gómez Ortega, J., Aguilera Puerto, D., Gámez García, J., 2011. Current situation and future perspectives on virgin olive oil elaboration process control. Rev. Iberoam. Autom. Inf. Ind. 8, 258–269.

CE, 2003. Regulation 1989/2003 of 6 November 2003 amending Regulation (EEC) No 2568/91 on the characteristics of olive oil-pomace oil and on the relevant methods of analysis. Technical Report. European Communities.

Christy, A.A., Kasemsumran, S., Du, Y., Ozaki, Y., 2004. The detection and quantification of adulterant in olive oil by near-infrared spectroscopy and chemometrics. Anal. Sci. 20, 935–940.

Cozzolino, D., Murray, I., Chree, A., Scaife, J.R., 2005. Multivariate determination of free fatty acids and moisture in fish oils by partial least-squares regression and near-infrared spectroscopy. LWT-Food Sci. Technol. 38, 821–828.

Eiben, A., Smith, J., 2003. Introduction to Evolutionary Computing. Springer.

El-Abassy, R., Donfack, P., Materny, A., 2009. Rapid determination of free fatty acid in extra virgin olive oil by raman spectroscopy and multivariate analysis. J. Am. Oil. Chem. Soc. 86, 507–511.

Huang, M., Zhu, Q., Wang, B., Lu, R., 2012. Analysis of hyperspectral scattering images using locally linear embedding algorithm for apple mealiness classification. Comput. Electron. Agric. 89, 175–181.
ImSpector, 2003. ImSpector Imaging Spectrograph User Manual. Spectral Imaging, Ltd.
Inarejos García, A.M., Gómez Alonso, S., Fregapane, G., Salvador, M.D., 2013. Evaluation of minor components, sensory characteristics and quality of virgin olive oil by near infrared (NIR) spectroscopy. Food Res. Int. 50.
Jamshidi, B., Minaei, S., Mohajerani, E., Ghassemian, H., 2012. Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of valencia oranges. Comput. Electron. Agric. 85, 64–69.
Jimenez, A., Beltran, G., Aguilera, M., Uceda, M., 2008. A sensor-software based on artificial neural network for the optimization of olive oil elaboration process. Sens. Actuat. B: Chem. 129, 985–990.
Latorre Carmona, P., Sotoca, J.M., Pla, F., 2012. Filter-type variable selection based on information measures for regression tasks. Entropy 14, 323–343.
Mailer, R.J., 2004. Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. J. Am. Oil. Chem. Soc. 81.
Marquez, A.J., 2003. Monitoring carotenoid and chlorophyll pigments in virgin olive oil by visible-near infrared transmittance spectroscopy. On-line application. J. Near Infrared Spectrosc. 11.
MathWorks, 2004. Genetic Algorithm and Direct Search Toolbox. The MathWorks.
MathWorks, 2013. Statistics Toolbox User's Guide. The MathWorks.
Monteiro, S.T., Kosugi, Y., 2007. Particle swarms for feature extraction of hyperspectral data. IECE Trans. Inf. Syst. E90D, 1038–1046.
Ruiz Altisent, M., Ruiz Garcia, L., Moreda, G., Lu, R., Hernandez Sanchez, N., Correa, E., Diezma, B., Nicola, B., Garca Ramos, J., 2010. Sensors for product characterization and quality of specialty crops. Comput. Electron. Agric. 74, 176–194.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.
Sinelli, N., Casale, M., Di Egidio, V., Oliveri, P., Bassi, D., Tura, D., Casiraghi, E., 2010. Varietal discrimination of extra virgin olive oils by near and mid infrared spectroscopy. Food Res. Int. 43, 2126–2131.
Suphamitmongkol, W., Nie, G., Liu, R., Kasemsumran, S., Shi, Y., 2013. An alternative approach for the classification of orange varieties based on near infrared spectroscopy. Comput. Electron. Agric. 91, 87–93.
Williams, P., Norris, K., 2004. Near infrared technology in the agriculture and food industries. Eds. Am. Cereal Assoc. Cereal Chem., chapter 8
Yao, H., Lewis, D., Sun, P., 2010. Spectral preprocessing and calibration techniques. In: Hyperspectral Imaging for Food Quality Analysis and Control. Academic Press, San Diego.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320.