Chemometric Characterization of Sunflower Seeds

Gastón Lancelle Monferrere, Silvana Mariela Azcarate, Miguel Ángel Cantarelli, Israel German Funes, and José Manuel Camiña

Abstract: The spectroscopic characterization of different varieties of sunflower seeds based on their oleic acid content is proposed. One hundred fifty samples of sunflower seeds from different places of Argentina were analyzed by near-infrared diffuse reflectance spectroscopy (NIRDRS). Seed samples were grounded and sieved without chemical treatment previous to the analysis. For the characterization, the used multivariate methods were: principal component analysis (PCA), cluster analysis (CA), linear discriminant analysis (LDA), and partial least square discriminant analysis (PLS-DA). By using PCA, CA, and LDA, and from the point of view of varieties of sunflower seeds, 2 groups were differentiated, based on the concentration of oleic acid: a low oleic group, which ranged from 15% to 25% w/w oleic acid; and the other one (mid-high oleic varieties) which ranged from 26% to 90% w/w oleic acid. However, by using the PLS-DA, 3 groups were correctly differentiated based on the concentration of oleic acid: low oleic (from 15% to 25% w/w oleic acid); mid oleic (26% to 76% w/w oleic acid); and high oleic ($\geq$ than 77% w/w oleic acid), demonstrating the high classification ability of this method. This multivariate characterization of sunflower seed varieties did not require chromatographic analysis to generate the matrix of concentrations, and only direct measures of NIRDRS spectra were required. This characterization can be useful to quickly know the variety of sunflower seed in the grain market.

Keywords: chemometric, infrared spectroscopy, oleic acid, seeds, sunflower

Practical Applications: This manuscript describes a method to determine 3 varieties of sunflower seeds (high, mid, and low oleic) The advantage of this method is to avoid the use of techniques that require long-time analysis.

Introduction

The sunflower plant (*Helianthus annuus L.*) is an important species widely used as crop around the world. Sunflower is an American crop that was first known in Europe after the European exploration journeys around the 1500s, expanding its use worldwide. In 2009, the total world production of sunflower seed was 3.24×10^7 ton, with a total harvest area of 2.37×10^7 ha. The world top ranking production in 2009 was Russia (6.45×10^6 ton), Ukraine (6.36×10^6 ton), and Argentina (2.50×10^6 ton). For these countries, the sunflower harvest areas were: Russia 5.59×10^6 ha, Ukraine 4.19×10^6 ha, and Argentina 1.82×10^6 ha (Food and Agriculture Organization of the United Nations 2011).

In 2009, the use of sunflower byproducts around the world, such as the production of sunflower oil was 1.32×10^7 ton; the first places were occupied by Russia with 2.80×10^6 ton, Ukraine 2.79×10^6 ton, and Argentina 1.41×10^6 ton. The distribution of sunflower oil production by continents was the following: Europe 8.79×10^6 ton; America 1.89×10^6 ton (South America 1.58×10^6 ton and North America 3.04×10^5 ton); Asia 2.03×10^6 ton; Africa 4.77×10^5 ton; and Oceania 2.44×10^4 ton (Food and Agriculture Organization of the United Nations 2011).

From the point of view of varieties, sunflower is frequently classified as low oleic (from 15% to 25% w/w oleic acid); mid oleic (26% to 76% w/w oleic acid); and high oleic ($\geq$ than 77% w/w oleic acid). Due to their oleic acid content, the high and mid oleic sunflower varieties are very much requested by edible oil factories to obtain oils with high oleic acid concentration, which have a higher commercial price in comparison with the rest of the common edible oils. On the other hand, there exists around the world a great consumption of sunflower oil used in the production of biodiesel, which involves the use of vegetable and animal oils, including sunflower varieties (Vicente and others 2006).

For the determination of sunflower varieties, it is necessary to determine the concentration of oleic acid in seeds by gas chromatography (GC), which is the most classical method for this determination (Hajimahmoodi and others 2005). The GC method requires an important time for sample preparation and chromatographic times, being this topic crucial for a fast commercialization of seeds for biodiesel production and edible oil factories.

On the other hand, the most frequently used multivariate tools applied to the classification of samples are: principal component analysis (PCA), cluster analysis (CA), linear discriminant analysis (LDA) (Massart and others 1997), and more recently, partial least square discriminant analysis (PLS-DA) (Schievano and others 2010; Aguilar and others 2011). PCA and CA are unsupervised multivariate methods, while LDA and PLS-DA are supervised methods (Fernandez 2005; Schievano and others 2010; Aguilar and others 2011). These 4 multivariate tools are a

MS 20120624 Submitted 5/3/2012, Accepted 6/18/2012. *Author Lancelle Monferrere is with Facultad de Ciencias Exactas, Naturales y Agrimensura, Univ. Nacional del Nordeste, Av. Libertad 5450 (3400) Corrientes, Argentina. Authors Azcarate, Cantarelli, and Camiña are with Facultad de Ciencias Exactas y Naturales, Univ. Nacional de La Pampa, Av. Uruguay 151 (6300) Santa Rosa, La Pampa, Argentina; and Inst. de las Ciencias de la Tierra y Ambientales, Av. Uruguay 151 (6300) Santa Rosa, La Pampa, Argentina. Author Funes is with Laboratorio de Control de Calidad, Gente de La Pampa SA, Ruta 1 km 171.5 (6330) Catriló, La Pampa, Argentina. Direct inquiries to author Camiña (E-mail: jcaminia@gmail.com, jcaminia@exactas.unlpam.edu.ar).*

doi: 10.1111/j.1750-3841.2012.02881.x

powerful combination that allows reaching new and interesting results.

Multivariate methods have been widely reported for the classification of edible oils, including the classification of genetic varieties of olive oil by HPLC-MS (Lerma-García and others 2009); infrared spectroscopy for the authentication of olive oils (Lerma-García and others 2010) and their geographical origin (Galtier and others 2007; Galtier and others 2008); the voltammetric method for the classification of edible oils (Gambarra-Neto and others 2009); classification of vegetable oils from their fatty acid content (Brodnjak-Voncina and others 2005); NMR spectroscopy, for the classification of edible oils and detection in olive oil adulteration, (Vigli and others 2003); classification of vegetable oils by GC-MS (Jakab and others 2002); thermal degradation of edible oils (Moros and others 2009); adulteration of olive oil by Visible and NIR spectroscopy (Downey and others 2002); determination of linoleic acid by attenuated total reflectance Fourier transform infrared spectroscopy (Kadamne and others 2009); and classification of edible oil using phosphorescence data (Arancibia and others 2008). However, there exist only a few reports that characterize plants and seeds by multivariate data analysis, including pumpkin seeds (Saucedo-Herna and others 2011), amaranth seeds (Aguilar and others 2011), onion (Galdón and others 2010), maize kernel (Williams and others 2009), green coffee (Alonso-Salces and others 2009), perilla seeds (Kim and others 2007), corn kernel (Weinstock and others 2006), and soybean (Roberts and others 2006).

For this reason, this article discusses the classification of 3 varieties of sunflower seeds based on their oleic content: low oleic (≤ 25% w/w oleic acid), mid oleic (between 26% and 76% w/w), and high oleic varieties (≥77% w/w oleic acid) using near-infrared diffuse reflectance spectroscopy (NIRDRS) (Hao and others 2009) and multivariate data analysis by PCA, CA, LDA, and PLS-DA. Spectral data consisted in the first derivative absorbance values from 1608 to 1708 nm, every 2 nm. The same variables were used for the 4 multivariate methods, obtaining, in all cases, a successful classification.

Materials and Methods

Instrumental

Diffuse reflectance infrared measurements were taken by a Brimrose NIRS model Luminar (Mass., U.S.A.) with acoustic-optic tuning filter (AOTF) and a rotator cup as sample cell.

For the confirmation of sunflower seed varieties (low, middle, and high oleic varieties), the quantification of oleic acid was carried out by chromatographic analysis, using a Varian Gas Chromatograph model GC 3900 (Calif., U.S.A.) with a Varian flame ionization detector (FID). A capillary column Varian factor FOUR VF-23 (cyanopropyl stationary phase, 30 m, 0.25 mm ID) was used. Analytical balance Ohaus model Pioneer (N.J., U.S.A.), Jack hydraulic press (Buenos Aires, Argentina), and Dalvo grinder model MCI (Buenos Aires, Argentina) were used.

The Unscrambler 6.11 software (CAMO-ASA, Trondheim, Norway) was used for the PCA and PLS-DA modeling, while CA and LDA were calculated using the Infostat software (Córdoba, Argentina).

Sampling and sample treatment

One hundred fifty samples (72 low oleic, 52 mid oleic, and 26 high oleic) varieties of sunflower seeds were obtained from different places of Argentina and harvested from 2009 to 2010. Samples were selected by an expert botanist, covering the maximum range of oleic acid in sunflower seeds.

To confirm the varieties in all samples, chromatographic analysis was carried out following the ISO 5590 method for the determination of fatty acids methyl esters by GC (International Organization for Standardization 1978; Milinsk and others 2008).

For the NIRDRS analysis, whole sunflower seeds samples, free from dust, were ground in a grinder. The fraction ≤2 mm was obtained by passing through a 2 mm sieve and placing in the sample rotator cup.

NIRDRS

A total of 20.00 g of whole ground and sifted sample seeds were placed in the rotator cup and the NIRDRS spectra were obtained. The total spectral range went from 1100 to 2200 nm every 2 nm. The sample rotator cup with the samples was irradiated with NIR radiation, and detected by acousto-optic tunable filter (AOTF). Twenty scans were obtained for every seeds sample and the total number of scans collected from every sample was averaged into a single spectrum. Measurements were carried out at room temperature.

Results and Discussion

PCA

PCA is a multivariate tool that can process an enormous amount of data produced by computers and other measurement techniques. PCA was used to search for data trends, combining the original variables. The PCA model was built using 50 variables, corresponding to the first derivative absorbance values, from 1608 to 1708 nm (Figure 1). To obtain this model, the validation method used was cross-validation. Using the selected variables, the model was obtained using only 3 principal components, which explain the 99.4% of the original information, allowing the building of a fit model. Figure 2 shows the classification obtained by PCA, through the scores plot. This figure shows 2 ellipses, grouping 2 sunflower varieties: mid-high oleic (upper ellipse)—*M-H* group—and low oleic (lower ellipse)—*L* group. In the *M-H* group, there were included 78 samples, while the *L* group had 72 samples.

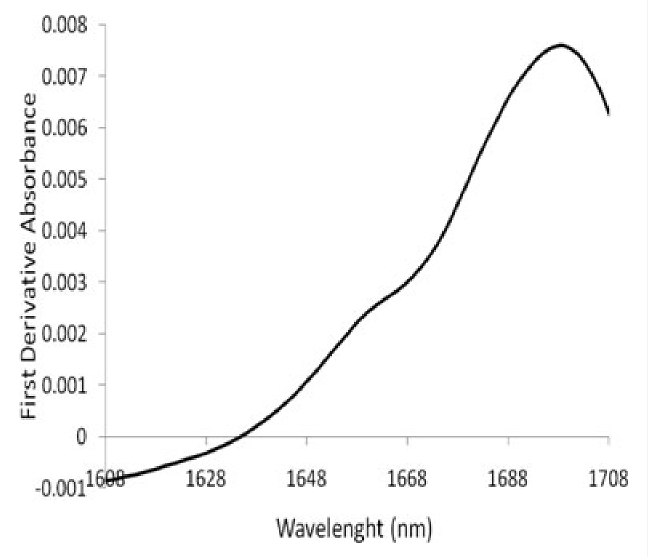

Figure 1–First derivative absorbance spectrum showing the wavelength range used to obtain the PCA model.

CA

As the PCA model, CA is used to classify objects, characterized by the values of a set of variables, into groups. It is therefore an alternative to PCA for describing the structure of a data table (Massart and others 1997). Due to its unsupervised nature, CA is frequently used to screen data for the clustering of samples. CA involves techniques that produce classification from unclassified data, which allows obtaining groups based on their similitude (Camiña and others 2008).

CA was carried out using the same data matrix that was used for PCA (150 samples and 50 variables). The complete linkage was used as a hierarchical linkage criterion of amalgamation that calculates the distances between objects of cluster, while the Euclidean distance was used as association criterion (Fernandez 2005). Figure 3 shows the dendogram plot for 150 samples of sunflower seeds. Again, the classification was successful, obtaining 2 groups with the same samples obtained through PCA. The left group corresponds to the *L* group (72 samples), while the right one corresponds to *M-H* group (78 samples).

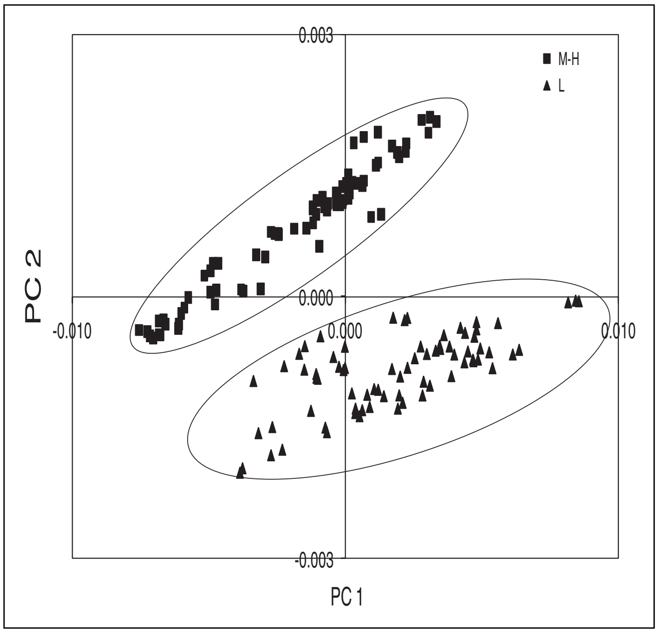

Figure 2–Score plot for the classification of mid-high (*M-H*) and low (*L*) sunflower seed varieties.

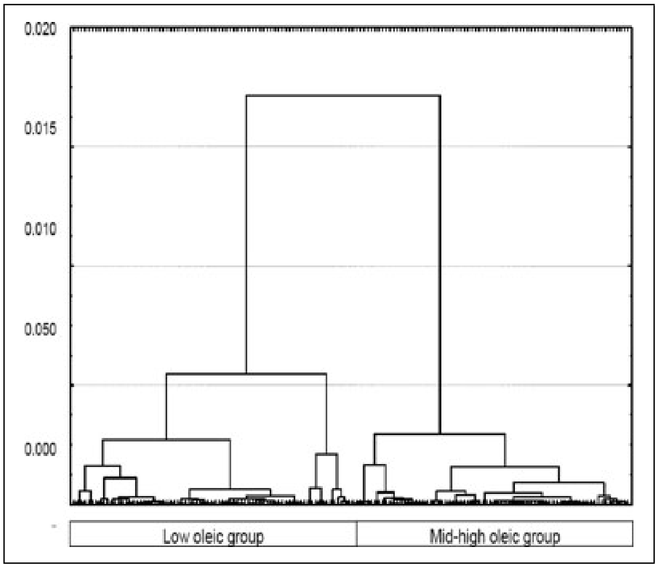

Figure 3–Dendogram of sunflower varieties by CA.

LDA

LDA is a supervised method used to find a theoretical value resulting in the best possible discrimination between *a priori* established groups. Discrimination is based on weighing the theoretical values for each variable in such a way as to maximize between-group variance in comparison to within-group variance (Johnson 2000). Discriminant analysis models involve sets of equations that are linear combinations of the independent variables, resulting in the maximum possible separation between groups (Hernandez and others 2005); these equations are known as discriminant function (Fernandez 2005).

The LDA model was constructed using cross classification criteria, using the original variables that were used in PCA and CA. One hundred four samples were used to build the model (calibration or training step), while 46 samples were used to validate the model (validation or prediction step). Table 1 shows the results of the LDA model, evidencing good fitting. The 46 samples were adequately classified; no sample was misclassified in the calibration and validation steps, which indicates that the LDA model has a prediction ability of 100%. On the other hand, the same classification results obtained for PCA and CA: *H-M* group with 78 samples and *L* group with 72 samples were obtained.

PLS-DA

PLS-DA is a supervised method that works building a PLS multivariate calibration model by means of partial least square regression, which is then used to predict unknown samples. As LDA, when samples are known, it is possible to obtain the degree of fit of a model and it offers the possibility to obtain a graphical classification of sunflower seeds by means of a PLS score plot. PLS-DA is an important multivariate tool, which was used as calibration tool some years ago. More recently, PLS has had new applications, for example, as a classification tool, by means of PLS-DA, from which a supervised model is obtained to classify groups from different categories.

For the selection of variables, the PCA served as a variable preselecting tool (Hernandez and others 2005). The PLS model was obtained using 3 principal components, which had 99.6% of the original information in the calibration step. To build the model, a cross-validation method was used, in which one sample is left out to build the model, and then this sample is used to obtain the prediction. When all samples are left out, the root mean square

Table 1–Results of the classification ability of the LDA model for mid-high (*M-H*[a]) and low (*L*) sunflower varieties.

	Prediction		
Data set	*M-H*[a]	*L*[b]	% Error
Calibration			
M-H[a]	53	0	0
L	0	51	0
Prediction			
M-H[a]	25	0	0
L	0	21	0
Total	78	72	0

[a] Mid-high oleic.
[b] Low oleic.

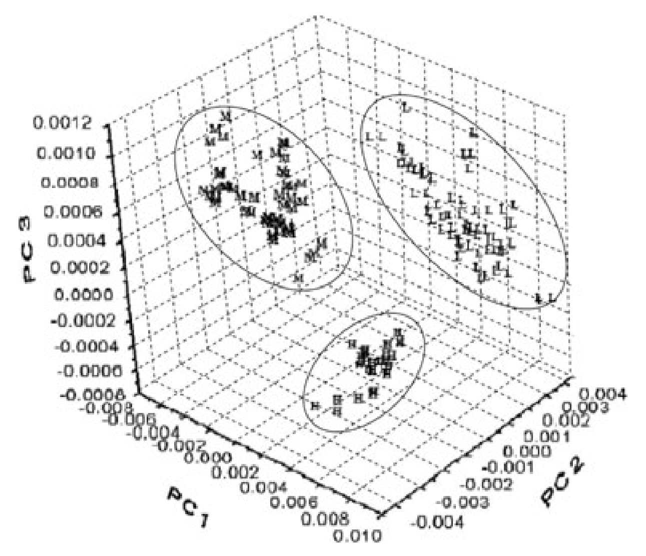

Figure 4–PLS-DA 3D score plot for 150 samples of sunflower seed varieties: high (*H*), mid (*M*), and low (*L*).

Table 2–Results obtained by PLS-DA.

Calibration	*H*[a]	*M*[b]	*L*[c]	Total	Correct (%)
H	20	0	0	20	100
M	0	20	0	20	100
L	0	0	20	20	100
Validation	*H*[a]	*L*[b]	*L*[c]	Total	Correct (%)
H	5	0	0	5	100
M	0	5	0	5	100
L	0	0	5	5	100

[a] High oleic.
[b] Mid oleic.
[c] Low oleic.

of error prediction (RMSEP) can be obtained, the value of which was low (0.18) for this model.

In Figure 4, the 3D classification score plot obtained with this model can be seen, where the 3 varieties of sunflowers seeds were distinguished: *H* (high oleic, 26 samples), *M* (mid oleic, 52 samples), and *L* (low oleic, 72 samples). After that, a PLS-DA table was obtained by using 20 random samples of sunflowers seeds in the calibration set and 5 samples used in the test set. Table 2 shows the results of the PLS-DA classification, where 100% of correct classification was obtained in the calibration (training) and validation (prediction) test sets.

Conclusions

This work has shown the ability of multivariate methods for classification of sunflower seeds. PCA, CA, and LDA provided a correct classification of 2 groups: low and mid-high oleic varieties. However, the PLS-DA supervised method showed a much better ability for graphical and discriminant analysis, due to its ability to recognize 3 separate groups based on the oleic acid concentration. On the basis of the results shown and in comparison with the other multivariate methods, PLS-DA was the best tool for the classification of sunflower seeds in low, mid, and high oleic types. The advantage of the proposed methods compared to previous works by NIRDRS-PLS is that the multivariate classification of varieties does not require chromatographic analysis to generate the matrix of concentrations, so getting the IR spectra can result in fast and economical classification of varieties of sunflower seeds. Due to the importance of the world sunflower seeds market, which determines the prices on the basis of the oleic acid concentration, this work can be useful for routine food laboratories and sunflower oil factories as a fast improvement in the classification of sunflower seeds.

Acknowledgments

The authors wish to thank the financial support from Univ. Nacional de San Luis, Univ. Nacional de La Pampa, and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Also, we want to thank to Prof. Roberto Olsina (Univ. Nacional de San Luis) for providing the Unscrambler 6.11 software; and Facultad de Agronomía, Univ. Nacional de La Pampa, for providing the InfoStat software.

References

Aguilar EG, Cantarelli MA, Marchevsky EJ, Escudero NL, Camiña JM. 2011. Multielemental analysis and classification of amaranth seeds according to their botanical origin. J Agric Food Chem 59:9059–64.

Alonso-Salces RM, Serra F, Remero F, Heberger K. 2009. Botanical and geographical characterization of green coffee (coffea arabica and coffea canephora): chemometric evaluation of phenolic and methylxanthine contents. J Agric Food Chem 57:4224–35.

Arancibia JA, Boschetti CE, Olivieri AC, Escandar GM. 2008. Screening of oil samples on the basis of excitation-emission room-temperature phosphorescence data and multiway chemometric techniques. Introducing the second-order advantage in a classification study. Anal Chem 80:2789–98.

Brodnjak-Voncina D, Cencic Kodba Z, Novic M. 2005. Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. Chemometr Intell Lab 75:31–43.

Camiña JM, Cantarelli MA, Lozano VA, Boeris MS, Irimia ME, Gil RA, Marchevsky EJ. 2008. Chemometric tools for the characterization of honey produced in La Pampa, Argentina, from their elemental content, using inductively coupled plasma optical emission spectrometry (ICP-OES). J Apicult Res 47:102–7.

Downey G, Mc Intyre P, Davies AN. 2002. Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the eastern mediterranean by visible and near-infrared spectroscopy. J Agric Food Chem 50:5520–5.

Fernandez CM. 2005. Quimiometría. Valencia: Publicaciones Univ. De Valencia. p. 247–52.

Food and Agriculture Organization of the United Nations (FAO). 2011. Faostat Database, [Accessed 2011 Nov 11]. Available From: Http://Faostat.Fao.Org/.

Galdón BR, Peña-Méndez E, Havel J, Rodríguez EMR, Romero CD. 2010. Cluster analysis and artificial neural networks multivariate classification of onion varieties. J Agric Food Chem 58:11435–40.

Galtier O, Dupuy N, Le Dréau Y, Ollivier D, Pinatel C, Kister J, Artaud J. 2007. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of nir spectra. Anal Chim Acta 595:136–44.

Galtier O, Le Dréau Y, Ollivier D, Kister J, Artaud J, Dupuy N. 2008. Lipid compositions and French registered designations of origins of virgin olive oils predicted by chemometric analysis of mid-infrared spectra. Appl Spectrosc 62:583–90.

Gambarra-Neto F, Marino GF, Araújo MCU, Galvão RKH, Pontes MJC, De Medeiros EP, Lima RS. 2009. Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis. Talanta 77:1660–6.

Hao Y, Cai W, Shao X. 2009. A strategy for enhancing the quantitative determination ability of the diffuse reflectance near-infrared spectroscopy. Spectrochim Acta A 72:115–9.

Hajimahmoodi M, Heyden YV, Sadeghi N, Oveisi MRJB, Shahbazian S. 2005. Gas-chromatographic fatty-acid fingerprints and partial least squares modeling as a basis for the simultaneous determination of edible oil mixtures. Talanta 66:1108–16.

Hernandez OM, Fraga JMG, Jimenez AI, Jimenez F, Arias JJ. 2005. Characterization of honey from the canary islands: determination of mineral content by atomic absorption spectrophotometry. Food Chem 93:449–58.

International Organization for Standardization (ISO). 1978. International Standard ISO No. 5509, p. 1.

Jakab A, Nagy K, Héberger K, Vékey K, Forgács E. 2002. Differentiation of vegetable oils by mass spectrometry combined with statistical analysis. Rapid Commun Mass Spectrom 16:2291–7.

Johnson DE. 2000. Métodos multivariados aplicados al análisis de datos. Mexico: Internacional Thompson Ed. p. 327–32.

Kadamne JV, Jain VP, Saleh M, Proctor A. 2009. Measurement of nonjugated linoleic acid (cla) in cla-rich soy oil by attenuated total reflectance-Fourier transform infrared spectroscopy (ATR-FTIR). J Agric Food Chem 57:10483–8.

Kim KS, Park SH, Choung MG. 2007. Nondestructive determination of oil content and fatty acid composition in perilla seeds by near-infrared spectroscopy. J Agric Food Chem 57:1679–85.

Lerma-García MJ, Concha-Herrera V, Herrero-Martínez JM, Simó-Alfonso EF. 2009. Classification of extra virgin olive oils produced at la comunitat valenciana according to their genetic variety using sterol profiles established by high-performance liquid chromatography with mass spectrometry detection. J Agric Food Chem 57:10512–7.

Lerma-García MJ, Ramis-Ramos G, Herrero-Martínez JM, Simó-Alfonso EF. 2010. Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy. Food Chem 118:78–83.

Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J.1997. Handbook of chemometrics and qualimetrics, Vol. B. Amsterdam: Elsevier. p. 57–68.

Milinsk MC, Matsushita M, Visentainer JV, De Oliveira CC, De Souza NE. 2008. Comparative analysis of eight esterification methods in the quantitative determination of vegetable oil fatty acid methyl esters (FAME). J Brazil Chem Soc 19:1475–83.

Moros J, Roth M, Garrigues S, De La Guardia M. 2009. Preliminary studies about thermal degradation of edible oils through attenuated total reflectance mid-infrared spectrometry. Food Chem 114:1529–36.

Roberts CA, Ren C, Beuselinck PR, Benedict HR, Bilyeu K. 2006. Fatty acid profiling of soybean cotyledons by near-infrared spectroscopy. Appl Spectrosc 60:1328–33.

Saucedo-Herna F, Lerma-García MJ, Herrero-Martínez, JM, Ramis-Ramos G, Jorge-Rodríguez E, Simó-Alfonso EF. 2011. Classification of pumpkin seed oils according to their species and genetic variety by attenuated total reflection Fourier-transform infrared spectroscopy. J Agric Food Chem 59:4125–9.

Schievano E, Peggion E, Mammi S. 2010. ^{1}H nuclear magnetic resonance spectra of chloroform extracts of honey for chemometric determination of its botanical origin. J Agric Food Chem 58:57–65.

Vicente G, Martínez M, Aracil JA. 2006. Comparative study of vegetable oils for biodiesel production in Spain. Energy Fuel 20:394–8.

Vigli G, Philippidis A, Spyros A, Dais P. 2003. Classification of edible oils by employing ^{31}P and ^{1}H NMR spectroscopy in combination with multivariate statistical analysis. A proposal for the detection of seed oil adulteration in virgin olive oils. J Agric Food Chem 51:5715–22.

Weinstock BA, Janni J, Hagen L, Wrigth S. 2006. Prediction of oil and oleic acid concentrations in individual corn (Zea Mays L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. Appl Spectrosc 60:9–19.

Williams P, Geladi P, Fox G, Manley M. 2009. Maize kernel hardness classification by near infrared (NIR) hyperspectral imaging and multivariate data analysis. Anal Chim Acta 653:121–30.