

# Near-infrared spectroscopy is feasible to discriminate hazelnut cultivars

Elisabetta Stella,<sup>1</sup> Roberto Moscetti,<sup>1</sup> Letizia Carletti,<sup>1</sup> Giuseppina Menghini,<sup>1</sup> Francesco Fabrizi,<sup>1</sup> Valerio Cristofori,<sup>1</sup> Danilo Monarca,<sup>1</sup> Massimo Cecchini,<sup>1</sup> Riccardo Massantini<sup>1</sup>

<sup>1</sup>Department of Agriculture, Forest, Nature and Energy, University of Tuscia, Italy; <sup>2</sup>Department for Innovation in Biological, Agro-food and Forest system, University of Tuscia, Italy

## Abstract

The study demonstrated the feasibility of the near infrared (NIR) spectroscopy use for hazelnut-cultivar sorting. Hazelnut spectra were acquired from 600 fruit for each cultivar sample, two diffuse reflectance spectra were acquired from opposite sides of the same hazelnut. Spectral data were transformed into absorbance before the computations. A different variety of spectral pretreatments were applied to extract characteristics for the classification. An iterative Linear Discriminant Analysis (LDA) algorithm was used to select a relatively small set of variables to correctly classify samples. The optimal group of features selected for each test was analyzed using Partial Least Squares Discriminant Analysis (PLS-DA). The spectral region most frequently chosen was the 1980-2060 nm range, which corresponds to best differentiation performance for a total minimum error rate lower than 1.00%. This wavelength range is generally associated with stretching and bending of the N-H functional group of amino acids and proteins. The feasibility of using NIR Spectroscopy to distinguish different hazelnut cultivars was demonstrated.

## Introduction

Hazelnuts are one of the most important raw material for the confectionary and chocolate industries and they are generally marketed shelled. They are also used to add fiber, flavor and texture to various foods (Kibar and Ortzuk, 2009). Moreover, hazelnuts are a source of bioactive compounds that might be incorporated into new health-related products or serve as substitutes for synthetic ingredients (Contini *et al.*, 2008). Different uses of hazelnuts are being explored including the use of unripe fruit (Moscetti *et al.*, 2012).

According to qualitative standards of the confectionary industry, the main commodity-related characteristics of the fruits, in relation with the process and use, are: fruit morphology, moisture content, shape uniformity, percent kernel, flaw incidences and the peelability as a consequence of the roasting process.

Moreover, the confectionary industries and the markets require whole hazelnuts from specific cultivar to enhance the characteristic of the traditional product with the aim of releasing a high quality product (Massantini *et al.*, 2009).

The most suitable techniques for sorting and grading commodities are based on electrical properties, near-infrared (NIR) Spectroscopy, sound/noise/vibration, ultrasound nuclear magnetic resonance, X ray, volatile emission and other (Sun, 2010; Singh *et al.*, 2010; Wang *et al.*, 2011). NIR spectroscopy has a number of desirable qualities including minimal need for sample preparation, good sample penetration, and a wide range of applications. It is fast, easy to use, environmentally benign, and is highly suited to rapid on-line inspection (Pasquini, 2003; Wang *et al.*, 2010).

NIR spectroscopy was used to compare moisture and acidity of several varieties of hazelnuts (Bellincontro *et al.*, 2005). More recently Bellincontro *et al.* (2009) evaluated the quality of hazelnuts during the storage by NIR spectroscopy and finally Moscetti *et al.* (2013) identified hazelnuts flaw by Vis/NIR spectroscopy.

The aim of this paper was to demonstrate the feasibility of the NIR spectroscopy use for hazelnut-cultivar sorting of 'Tonda Gentile Romana' (TGR), 'Nocchione' (NOC) and 'Tonda di Giffoni' (GIF) that represent the main-grown varieties in Central Italy. The reason is to comply with the confectionary-industries' expectation that require special characteristics in relation with the specific use of the raw material itself and comply with the requirements of the regulation (G.U. n.186, 12-08-2009) that sets in 90% the maximum percentage of 'Tonda Gentile Romana' plus 'Nocchione' and 'Tonda di Giffoni'.

## Materials and methods

Whole hazelnut kernels (round shape cvs. 'Tonda Gentile Romana', 'Nocchione' and 'Tonda di Giffoni') were obtained from the Assofrutt

Correspondence: Riccardo Massantini, DIBAF, Tuscia University, Via S. Camillo de Lellis snc, 01100, Viterbo, Italy.

Tel. +39.0761.357496 - Fax: +39.0761.357498. E-mail: massanti@unitus.it

Key words: *Corylus avellana L.*, Near Infrared spectroscopy, hazelnut sorting

Contributions: the authors contributed equally.

Conflict of interests: the authors declare no potential conflict of interests.

Funding: this work was supported by the Ministry of Agricultural and Forestry Policy (MIPAAF ML.F.COL. 2012 – D.D. 17304).

Conference presentation: part of this paper was presented at the 10th Italian Conference AIA (Associazione Italiana di Ingegneria Agraria), 2013 September 8-12, Viterbo, Italy.

©Copyright E. Stella *et al.*, 2013

Licensee PAGEPress, Italy

Journal of Agricultural Engineering 2013; XLIV(s2):e55

doi:10.4081/jae.2013.s2.e55

This article is distributed under the terms of the Creative Commons Attribution Noncommercial License (by-nc 3.0) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

facility (Caprarola, Viterbo, Central Italy) in October 2012. The moisture content was, for all the samples,  $4.00 \pm 0.11\%$ . Hazelnut spectra were acquired from 600 fruit per each cultivar sample, using a Luminar 5030 Acousto-Optic Tunable Filter-Near Infrared (AOTF-NIR) Miniature “Hand-held” Analyzer (Brimrose Corp., Baltimore, USA). The instrument was equipped with a reflectance post-dispersive optical configuration and an indium gallium arsenide (InGaAs) array (range 1100-2300 nm, 2-nm resolution) with a scanning time of 60 ms. Each acquired spectrum was the average of 10 scans. The reference spectrum was automatically measured by the instrument as described by Cayuela and Weiland (2010). Two spectra were acquired from opposite sides of the same hazelnut. Diffuse reflectance spectra were acquired and converted to transmittance measurements using SNAP! 2.04 software (Brimrose Corp.). Transmittance spectral data were transformed into absorbance ( $A = \log T^{-1}$ ) using R 2.15.2 statistics software.

Features for use in classification were extracted from the spectra as raw absorbance values at each wavelength. Further, these features were extracted following a variety of spectral pretreatments including row-Mean Centering (rMC), Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and Savitzky-Golay first and second derivatives ( $df$ ,  $d^2f$ ) with second order polynomial (from 9 nm to 57 nm smoothing points with a step of 6 nm) (Savitzky and Golay, 1964; Boysworth and Booksh, 2008). High frequency noise was filtered from rMC, SNV, MSN and non-pretreated spectra by applying a Savitzky-Golay filter with 5-nm smoothing points. For each pre-treated and not-pre-treated dataset, Mean Centering (MC) was also tested, and a simulated annealing algorithm (SAA) was used to seek k-feature subsets (where k ranged from 2 to 6) which are optimal, as surrogates for the whole dataset (Cerdeira *et al.*, 2012). A maximum of 6-feature selection was computed to minimize the overfitting. An iterative Linear Discriminant Analysis (LDA) algorithm was used to choose a relatively small set of variables for the ability to correctly classify the samples. For each iteration of the algorithm, 50% of the samples were used as a training set, 25% were used for validation and 25% were used as test set. No outlier selection was computed.

The optimal group of features selected for each trial was analyzed using Partial Least Squares Discriminant Analysis (PLS-DA). For each combination of features a Receiver Operating Characteristics (ROC) graph was plotted for 20 instances (from 0.0 to 1.0 sound prior probability and from 1.0 to 0.0 unsound prior probability with a step of 0.05). The ROC curve is a two-dimensional depiction of classifier (combination of features) performance. It contains plots of the ‘false positive rate’ and the ‘true positive rate’ (also respectively called ‘1 – specificity’ and ‘sensitivity’) as a function of the threshold value (Fawcett, 2006):

$$\text{False positive rate} = 1 - \frac{\text{True negatives}}{\text{False positives} + \text{True negatives}} \quad (1)$$

$$\text{True positive rate} = \frac{\text{True negatives}}{\text{Total positives}} \quad (2)$$

The total error rate represents the percentage of hazelnuts in the validation set incorrectly classified. False positive (fp) errors occur when a good product is classified as bad, while false negative (fn) errors occur when a defective product is misclassified as good.

For each ROC plot the relative Area Under the ROC Curve (AUC) was additionally used to evaluate the performance of each computed linear discrimination function (Fawcett, 2006), because it is statistically consistent and more discriminating than the total error rate, as demonstrated by Ling *et al.* (2003). Consequently, this makes it possible to obtain different AUCs for linear discrimination functions providing the same error rate for a given threshold value. The AUC takes on values between 0.5 and 1.0. An AUC value close to 0.5 indicates that the two groups are not distinctly different and therefore the classifiers have a weak discriminative ability. An AUC value close to 1.0 indicates that the features have strong discriminative power and a low tendency for overlap in the distributions of the group scores (Luo *et al.*, 2012). The AUC value corresponds to the percentage of time a random selection from the positive group will produce a score greater than a random selection from the negative group (Fielding and Bell, 1997). For this research, an AUC of 0.9 was used as a threshold for acceptable discriminative performance. Thus, classifiers with an AUC equal to or less than 0.9 were discarded.

Statistical pretreatments and analysis, and data normalizations were performed using R 2.15.1 statistics software in combination with TISEAN 3.0.1 package (Hegger *et al.*, 2012), and MASS 7.3-21, hyperspec 0.98, mda 0.4-2, PLS 2.3-0 and RTISEAN 3.0.14 R-packages (CRAN, 2012).

## Results and discussions

There are no scientific studies on NIR spectroscopy to recognize different hazelnut varieties and it is impossible to find data for wavelength selection related with hazelnuts discrimination. Spectra included in the range 1100-2300 nm were acquired because this range was positively used in other studies on vegetable oil characterization (Yildiz *et al.*, 2001), representing the main constituent of hazelnuts with a content above 60% on dry weight (Cristofori *et al.*, 2008; USDA, 2009). Consequently the knowledge of chemical composition of the hazelnut and relative bibliography informs of the relation between molecules and wavelength that are more frequently selected from statistical models during the experimental phase. To recognize the three varieties (GIF, NOC, TGR) a predictive model LDA was tested. However, the

Test #	Dataset				Features (nm)						Error rates (%)			AUC (%)
	Scatt. C.	Norm.	SG*	Deriv.	1	2	3	4	5	6	fp	fn	total	
01	MSC	-	45	df	1136	1822	1848	2026	-	-	0.00	0.02	0.01	99.98
02	-	-	45	-	2000	2060	-	-	-	-	0.00	0.02	0.01	99.94
03	-	-	53	-	1360	1888	2000	2058	2182	-	0.00	0.02	0.01	99.89
04	MSC	-	57	df	1142	1552	1726	1872	2034	-	0.00	0.02	0.01	99.87
05	-	-	13	-	1518	2016	2050	2174	-	-	0.00	0.02	0.01	99.81
06	-	MC	21	-	1990	2058	2136	-	-	-	0.00	0.02	0.01	99.68

obtained discriminating performances are not suitable because they showed a high classification error for 'Nocchione' and 'Tonda Gentile Romana' cultivars. Therefore, the initial purpose for developing a 3-classes discriminant function was discarded. Consequently, the spectral data from the NOC and the TGR samples were merged into a new class sample, corresponding to 'Nocciola Romana DOP' (NRM) and a new 2-classes discriminant model (GIF and NRM) was developed.

The spectral region most frequently selected by the 2-classes-iterative LDA was the 1980-2060 nm range, which corresponds to best discrimination performance for a total minimum error rate lower than 1.00%. This wavelength range is generally associated with stretching and bending of the N-H functional group of amino acids and proteins. Features from the aforementioned range were frequently coupled with the wavelengths included in the 2120-2180 nm range, which typically represent the vibrations of C-H/C=O groups of lipids and N-H/C=O groups of proteins and amino acids (Workman and Weyer, 2008).

Discriminant functions with excellent performances were also obtained using features selected from the 1150-1650 nm spectral region, which represents the first overtone of stretching vibrations of O-H group (water, free fatty acids and monoglycerids) and overtone and stretching of combination and deformation of the -CH group ( $\text{CH}_3$ -,  $-\text{CH}_2$ - and *cis*  $\text{R}_1\text{-CH=CHR}_2\text{CH}_3$ -) (Yildiz *et al.*, 2001; Christy *et al.*, 2004). In addition, other wavelengths close to 1800 nm, which are descriptive of the fatty acid composition (Bewig *et al.*, 1994), were also chosen by the iterative LDA function as features having good classification performances.

The most accurate classification of 0.00% fp, 2.00% fn and 1.00% total error was achieved using first derivative preprocessing combined with a Savitzky-Golay filter with 45 smoothing points and a Multiplicative Scatter Correction algorithm (Test #1) (Table 1). Features selected were *Abs*[1136 nm], *Abs*[1822 nm], *Abs*[1848 nm] and *Abs*[2026 nm]. For these features, the AUC value was 0.9999, corresponding to an excellent discrimination performance of the selected classifier. However, the amount of wavelengths required in an online sorting device would be conditional on the spectral pretreatment used. Obtaining scatter-corrected spectrum requiring to datasets of full spec-

tra. Consequently, Test #1 and other Tests performed with scatter correction techniques are not suitable for fast- and inexpensive-online-detection devices. With this in mind, the best classification resulted from a Savitzky-Golay filter with 45 smoothing point (Test #02), resulting in excellent discrimination between classes (0.9995 AUC): *Abs*[2000 nm] and *Abs*[2060 nm], which correspond to overtones related with amino acids and proteins composition of hazelnut (Figure 1). The PLS-DA of test #10 indicates that the 96.91% of the variance was explained by LV1 and the 3.09% of residual variance from LV2 was essential was a complete description of the samples (Figure 2).

These results need to be supported by chemical assessments to confirm the potential compositional differences observed by NIR spectroscopy and this would be a great reason to continue the experimentation.

## Conclusions

In the present study, the feasibility of using NIR Spectroscopy (from 1100 nm to 2300 nm) to discriminate different hazelnut cultivars (cvs. 'Tonda Gentile Romana', 'Nocchione' and 'Tonda di Giffoni') was demonstrated. Features for discriminant analysis that showed the best classification results were generally in the 1150-1650 nm, 1800 nm, 1980-2060 nm and 2120-2180 nm spectral bands. The 2 classes discriminant models showed a very high performance of the NIR Spectroscopy for a suitable classification of 'Nocciola di Giffoni IGP' and 'Nocciola Romana DOP'. The best classification performance (0% fp, 2% fn, 1% total error and 0.9995 AUC) were obtained using a Savitzky-Golay filter with 45-smoothing points, and *Abs* [2000 nm] and *Abs* [2060 nm] wavelengths.

Finally, the method used here allows reliable classification of hazelnuts cultivars based on a maximum of 6 wavelengths selected from within the spectral range from 1100-2300 nm, thus providing the means for a rated, on line detection system.

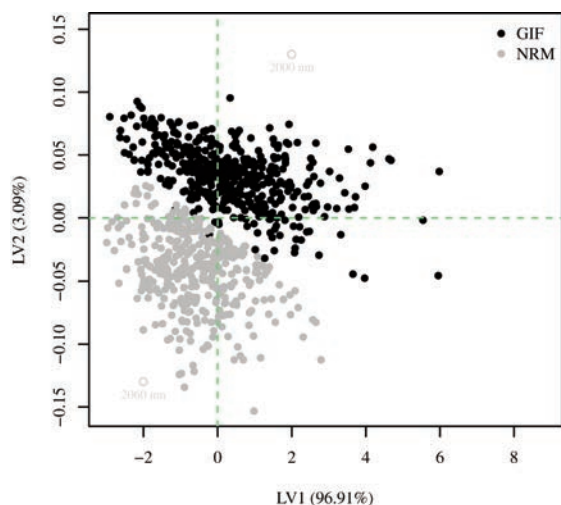


Figure 1. Classes spatial distribution along the LD1 dimension and relative overlaid density plot obtained from the three features selected from the spectra acquired from the flat side (fs) of the fruit. Spectra pretreated with the Savitzky-Golay filter (45-smoothing points).

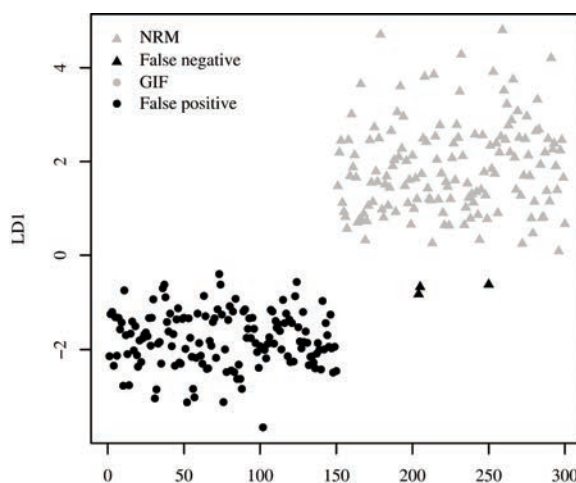


Figure 2. PLS-DA score plot of the two features selected during the Test #10 (Table 1). Dataset pretreated with the Savitzky-Golay filter (45-nm smoothing points). Percents of the explained variance are reported in

## References

- Bellincontro, A., Fracas, A., Anelli, G., Mencarelli, F., Di Natale, C., Esposito, G. Use of NIR Technique to measure the acidity and water content of hazelnuts. Proc. VI Intern. Congress on Hazelnut published Tous, J., Rovira, M. and Romero, A. Acta Horticult. 2005;686:499-503.
- Bellincontro, A., Mencarelli, F., Forniti, R., Valentini, M. Use of NIR-AOTF Spectroscopy and MRI for quality detection of whole hazelnuts. Proc.VII Intern. Congress on Hazelnut published Varvaro, L. and Franco, S. Acta Horticult. 2009;845:593-7.
- Bewig, K.M., Clarke, A.D., Roberts, C., Unklesbay, N. Discriminant analysis of vegetable oils by near-infrared reflectance spectroscopy. J. Am. Oil Chem. 1994;2:195-200.
- Boysworth, M.K., Booksh, K.S. Aspects of multivariate calibration applied to near-infrared spectroscopy. In Burn, D. A., & Ciurczak, W. E. (Eds.), Handbook of Near-Infrared Analysis (3rd ed.) 2008;207-29. CRC Press.
- Cayuela, J.A., Weiland, C. Intact orange quality prediction with two portable NIR spectrometers. Postharvest Biol. Tec. 2010;58:113-20.
- Cerdeira, J.O., Silva, P.D., Cadima, J., Minhoto, M. Package 'subselect' 0.12-2 for use with R. Selecting variable subsets. 2012.
- Christy, A.A., Kasemsumran, S., Du, Y., Ozaki, Y. The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics. Anal Sci 2004;20:935-40.
- Contini, M., Baccelloni, S., Massantini, R., Anelli, G. Extraction of natural antioxidants from hazelnut (*Corylus avellana* L.) Shell and skin wastes by long maceration at room temperature. Food Chem 2008;110:659-69.
- Cristofori, V., Ferramondo, S., Bertazza, G., Bignami, C. Nut and kernel traits and chemical composition of hazelnut (*Corylus avellana* L.) cultivars. J. Sci. Food Agr. 2008;88:1091-8.
- Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters 2006;27:861-74.
- Fielding, A.H., Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 1997;24:38-49.
- Italian Regulation. Disciplinare di produzione della denominazione di origine protetta 'Nocciola Romana' DOP. In: Official Journal No. 186, 12/8/2009.
- Hegger, R., Kantz, H., Schreiber, T. Nonlinear Time Series Analysis (TISEAN). 2012.
- Kibar, H., Öztürk T., The effect of moisture content on the physico-mechanical properties of some hazelnut varieties. J Stored Prod Res 2009;45:14-8.
- Ling, C.X., Huang, J., Zhang, H. AUC: a better measure than accuracy in comparing learning algorithms. Lect. Notes Artif. Int. 2003;2671:329-41.
- Luo, X., Takahashi, T., Kyo, K., Zhang, S. Wavelength selection in VIS/NIR spectra for detection of bruises on apples by ROC analysis. J Food Eng. 2012;109:457-66.
- Massantini, R., Moschetti, R., Monarca, D., Cecchini, M., Contini, M., Mordacchini Alfani, M.L. Influence of cover crop and double harvest on storage of fresh hazelnuts (*Corylus avellana* L.). Adv. Hortic. Sci 2009;23:231-7.
- Moschetti, R., Frangipane, M.T., Monarca, D., Cecchini, M., Massantini, R. Maintaining the quality of unripe, fresh hazelnuts through storage under modified atmospheres. Postharvest Biol. Tec. 2012;65:33-8.
- Moschetti, R., Haff, R.P., Aernouts, B., Saeys, W., Monarca, D., Cecchini, M., Massantini, R. Feasibility of Vis/NIR spectroscopy for detection of flaws in hazelnut kernels. J Food Eng 2013;118:1-7.
- Pasquini, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. 2003;14:198-219.
- Savitzky, A., Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 1964;36:1627-39.
- Singh, C.B., Jayas, D.S., Paliwal, J., White, N.D.G. Identification of insect-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. Comput. Electron. Agr. 2010;73:118-25.
- Sun, D. Hyperspectral imaging for food quality analysis and control. Academic Press, 2010. London.
- United States Department of Agriculture (USDA). National nutrient database for standard reference - Release 22, 2009. URL <http://ndb.nal.usda.gov/>
- Wang, J., Nakano, K., Ohashi, S., Takizawa, K., He, J.G. Comparison of different modes of visible and near-infrared spectroscopy for detecting internal insect infestation in jujubes. J. Food Eng. 2010;101:78-84.
- Wang, J., Nakano, K., Ohashi, S. Nondestructive detection of internal insect infestation in jujubes using visible and near-infrared spectroscopy. Postharvest Biol. Tech. 2011;59:272-9.
- Workman, J. Jr., Weyer, L. Practical Guide to Interpretive Near-Infrared Spectroscopy. CRC Press, 2008;51-69.
- Yildiz, G., Wehling, R.L., Cuppett, S.L. method for determining oxidation of vegetable oils by near-infrared spectroscopy. J. Am. Oil Chem. Soc. 2001;5:495-502.