Grape seeds characterization by NIR hyperspectral imaging

Francisco J. Rodríguez-Pulido[a]; Douglas F. Barbin[b]; Da-Wen Sun[b*]; Belén Gordillo[a]; M. Lourdes González-Miret[a]; Francisco J. Heredia[a].

[a] Food Colour & Quality Laboratory, Dept. Nutrition & Food Science. Facultad de Farmacia. Universidad de Sevilla. 41012-Sevilla, Spain.

[b] Food Refrigeration and Computerised Food Technology (FRCFT), School of Biosystems Engineering, University College Dublin, National University of Ireland, Agriculture & Food Science Centre, Belfield, Dublin 4, Ireland.

[*] Corresponding author. Tel: +353-1-7167342, Fax: +353-1-7167493, E-mail: dawen.sun@ucd.ie, Website: www.ucd.ie/refrig; www.ucd.ie/sun.

ABSTRACT

Phenolics compounds in grape seeds are responsible for numerous properties in wine and these compounds change during the whole development of grape. Currently, the moment of harvest is normally determined according to the sugar level in the pulp of grapes. Nonetheless, the stage of maturation in grape seeds should be taken into account more frequently to decide the appropriate harvest period. There are chemical and sensory analyses to assess the stage of maturation of grape seeds but they are destructive and time-consuming. The hyperspectral imaging arises as an alternative technology to characterize the grape seeds according to their chemical attributes, and the current work aimed to non-destructively characterize grape seeds in regard of the variety and stage of maturation. For this purpose, 56 samples of seeds from two red grape varieties (Tempranillo and Syrah) and one white variety (Zalema) in two kinds of soil were selected to assess their features based on the reflectance in the near-infrared (NIR) spectra by using prediction models (partial least squares regression) and multivariate analysis methods (principal component analysis and general discriminant analysis). In this study, a reliable methodology for predicting the stage of maturation was developed, and it was shown that it was possible to distinguish the variety of grape and even the type of soil from hypespectral images of grape seeds.

Keywords: hyperspectral imaging, *Vitis vinifera*, grape seeds, partial least squares regression, principal components analysis.

1. INTRODUCTION

Both genetic and environmental effects create significant variation in the amount of each component in grapes and their seeds. These compounds play an important role on both red and white wine. Structure, bitterness, astringency and body are some attributes given to wine by phenolic compounds in grape seeds (Escribano-Bailón et al. 2001; Gawel, 1998). The composition of the seeds changes along the maturation until the grapes reach the ripeness, affecting the sensory properties of wine. Chemical analyses are the most widely accepted reference methods for determining seeds composition. However, these methods frequently are destructive, requiring lengthy preparation procedures, and entails large samples of grain. Sensory analysis is another approach to assess the condition of the seeds. However, this technique is time consuming and it is difficult to be carried out in an objective manner (Rousseau and Delteil, 2000). Conventional near infrared spectroscopy has been already applied to grape seeds to determine their chemical composition. Nevertheless, spectroscopy by transmission requires lengthy and laborious preparation procedures as freeze-drying, grinding, and extraction processes are needed before sample presentation to the equipment (Ferrer-Gallego et al. 2010).

Computer vision systems have the limitation of only acting on the surface of materials, thus allowing only identification of external features. However, these phenolic compounds such as catechin and epicatechin are mainly concentrated within the outer layer of grape seeds (Thorngate and Singleton, 1994). Near-infrared (NIR) hyperspectral imaging is a powerful technique which has been used in a number of applications in agricultural products (Baye et al. 2006; Shahin and Symons, 2011). By combining both spatial and spectral features, this technique provides an alternative, non-destructive technology for measuring constituents of biological materials. Moreover,

this technique provides a way to distinguish among varieties in foodstuffs that would be very difficult by visual inspection (Choudhary et al. 2009). In hyperspectral imaging, the utilisation of NIR measurements allows the simultaneous determination of multiple constituents in a sample, since organic molecules have specific absorption patterns in the NIR region that can be used to characterize the chemical composition of the substance being examined (Williams and Norris, 2001).

Hyperspectral imaging provides a high spectral resolution allowing for models to utilize the spectral information in combination with the image data. This approach is useful in applications where spectral information may not be sufficient, where similar constituents need to be separated. NIR hyperspectral imaging systems acquire spectral images, also known as hypercubes, which are three-dimensional data matrix where the first two axes (x and y) of the matrix represent the spatial coordinates, while the third (λ) axis depicts the spectral dimension. Hundreds of single channel black and white (grayscale) images, where each image represents a single band of spectral wavelength, are stacked on top of each other to produce a hypercube (Burger and Geladi, 2006). NIR imaging thus has the potential application to early detection of chemical changes and can be superior to visual based systems.

Hyperspectral images (HSI) are large in size and spectral data are highly correlated, requiring the application of multivariate data analysis techniques for data exploration. As with NIR spectroscopy, chemometric techniques are applied to decompose the image dataset, pre-process and perform regression or classification analyses. Identifying most useful wavelengths is a good means of reducing the large amounts of data.

Several studies have illustrated the potential of using hyperspectral imaging based on NIR range to develop a model able to predict and classify barley kernels according to germination stage (Engelbrecht et al. 2010; Munck and Møller, 2004). Williams *et al.*

(2009) employed PLS discriminant analysis to classify maize kernels into hardness categories. Combination of the visual and NIR spectral ranges (400-1000 nm) were used to classify α-amylase activity into two classes and to detect sprout damage to Canada Western Red Spring wheat (Xing et al. 2010).

In NIR spectral measurements, physical characteristics of samples and variations in the instrument response can cause light scattering effects, originating spectral differences that are not correlated to the analysed responses. Once the spectral effect of light scatter is different from that of chemically based light absorption, scattering effects can be corrected in the data by a sort of mathematical treatment. Among the most commonly used methods for spectral correction are multiplicative scatter correction (MSC), standard normal variate (SNV) and derivation (Geladi et al. 1985; Isaksson and Næs, 1988; Kaihara et al. 2002; Pizarro et al. 2004; Windig et al. 2008).

MSC is a set-dependent method that corrects the scatter level of all spectra in a dataset for multiplicative (slope) and additive (offset) effects, with equivalent results to SNV. The main difference is that MSC uses the calculated mean spectrum of the dataset, while SNV standardises each spectrum using only the data from that particular spectrum (Barnes et al. 1989).

First derivative transforms are useful for eliminating baseline offset variations within a set of spectra, but the slope becomes a constant term. The 2^{nd} derivative can help separate overlapping peaks and sharpen spectral features, being a very effective method to eliminate both the baseline offset and slope from a spectrum. At the typical spectral sampling interval of hyperspectral systems, derivatives of the second order or higher should be relatively insensitive to variations in illumination (Nicolaï et al. 2007; Osborne et al. 1993). However, there is still no standard procedure to decide which

115 spectral pre-processing produce best results to a given dataset and often the only
116 approach is trial and error.
117 In this study, pre-processing techniques were implemented to analyze the impact of
118 systematic noise in the spectral data acquired from the grape seeds, while retaining
119 useful information for sample characterization. The functionality of the pre-processing
120 techniques was compared by the prediction ability of PLSR models relating the pre-
121 processed spectra and the harvest period regardless the varieties of grape. This step
122 aimed to establish a methodology that would allow a systematic approach for further
123 studies.
124 Partial least squares regression (PLSR) is a procedure used to relate a large number of
125 independent variables (predictors) to the prediction of one or few response variables
126 (observations). This technique is particularly effective in spectral analysis since it
127 reduces a great number of highly correlated original descriptors (wavelengths) to a new
128 variable space based on orthogonal factors called latent variables. In PLSR, the
129 information not contained in the factors is described by a residual value calculated for
130 each spectrum. The resulting model has the following form:
131 $$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \quad (1)$$
132 where $\boldsymbol{y}$ is the response matrix of independent instrumental and sensory measurements
133 (N samples × 1), $\boldsymbol{X}$ is the predictors matrix (N samples × W wavelengths), $\boldsymbol{b}$ is the
134 matrix of regression coefficients obtained from PLS analysis, and $\boldsymbol{e}$ is the matrix of
135 residual information not contained in the factors.
136 By small sample sizes, validated misclassification rates can be very sensitive to the
137 choice of segmentation; hence, it can be difficult to assess whether an obtained rate of
138 misclassification is substantially biased relative to random results (Kjeldahl and Bro,
139 2010). Hence, results of the leave-one-out cross-validation (LOOCV) and test set

approaches were compared to corroborate the prediction model results. Performance of the prediction models was evaluated using the root mean square error of calibration (RMSEC), the coefficient of determination in calibration and cross-validation (R^2_C and R^2_{CV}), the root mean square error of cross-validation (RMSECV), the prediction error sum of squares (PRESS) and number of latent variables required (LV) (Fawcett, 2006). The best model selected should have high coefficient of determination (R^2_C and R^2_{CV}), and low standard errors (SEC and SECV), in addition to a small difference between SEC and SECV (ElMasry et al. 2011). The optimal number of latent variables (LV) for establishing the calibration model was determined at the minimum value of predicted residual error sum of squares (PRESS) under cross-validation.

Multivariate analysis methods have the advantage of being able to deal with large complex co-linear spectral data and to reduce these data to a lower dimension without loss of useful information. A PCA was conducted to analyse the variance among the spectral information obtained from the samples. The matrix expression of the PCA model for the spectral data can be obtained below:

$$\boldsymbol{R} = \boldsymbol{SV} + \boldsymbol{E} \qquad (2)$$

where R is the spectral reflectance matrix (n × w) extracted from the corrected image; S is the score matrix (n × p); V is the eigenvector matrix (p × w); E is a residual matrix (n × w); n is the number of spectra; w is the number of wavelengths and p is the number of principal components (Park et al. 2001). The loadings resulting from PCA could also be used to identify and investigate the influence of the wavelengths (variables).

General discriminant analysis (GDA) is another commonly used technique for data classification and dimensionality reduction. This method maximizes the ratio between-class variance to the within-class variance in any particular dataset thereby guaranteeing maximal separability. This statistical technique requires a qualitative variable

(dependent variable) and at least two quantitative or dichotomous variables (independent variables). It is a method of classification whose aim is to estimate through linear functions (discriminant functions) of the independent variables (relative reflectance at multiple wavelengths) the probability that one of the cases belongs to each of the groups defined by the categories of the dependent variable (varieties of grape seeds) (González-Miret et al. 2006; Johnson, 2000).

The main objective of the present study was to investigate the potential of using NIR hyperspectral reflectance imaging technique as a fast and non-invasive method to characterize grape seeds according to varieties and stage of maturation. Specific objectives were to (1) establish a satisfactory approach to extract spectral data from hyperspectral images of grape seeds acquired in the NIR range (900-1700 nm), (2) study whether spectral pre-processing methods can improve robustness of the prediction models, (3) identify the most significant wavelengths linked to the seeds characteristics, (4), build robust PLSR models to relate spectral information and stage of maturation using selected wavelengths.

2. MATERIALS AND METHODS

2.1. Sample preparation

The vineyards sampled are included under the "Condado de Huelva" Designation of Origin, in Southwestern Spain, harvested in 2011. Two red varieties (Tempranillo and Syrah) and one autochthonous white variety (Zalema) cultivated in two kinds of soil (Sand and Clay) were used. According to availability in each variety in vineyards, samples were taken twice a week from late June until post-harvest on early September. Sampling was carried out by taking a pair of berries from alternate grapevines and from both sides up to reach 2 kg of berries, in order to ensure the representativeness of the

samples. Once in laboratory, one hundred berries were randomly taken and seeds were removed, dried, and frozen at −20 ºC until acquisition of hyperspectral images.

2.2. Acquisition and processing of hyperspectral images

NIR spectral images were acquired in the reflectance mode using a pushbroom hyperspectral imaging system (Figure 1). The system comprises a spectrograph (ImSpector, N17E, Spectral Imaging Ltd, Finland), a charged-couple device (CCD) camera with C-mount lens (Xeva 992, Xenics Infrared Solutions, Belgium), a translation stage (MSA15R-N, AMT-Linearways, SuperSlides&Bushes Corp., India) with two tungsten-halogen lamps (V-light, Lowel Light Inc, USA) providing illumination, data acquisition software (SpectralCube, Spectral Imaging Ltd., Finland), and a computer. The conveying translation stage was driven by a stepping motor (GPL-DZTSA-1000-X, Zolix Instrument Co, China) with a user-defined speed of 2.7 cm s^{-1} synchronized with the image acquisition by the data acquisition software. The horizontal axis of the spatial dimension of the images (x) is fixed, while the sample was scanned underneath the field of view (FOV) of the camera to determine the size of the vertical axis of the spatial dimension (y) of the spectral images. The conveyer speed was adjusted to fit the predetermined camera exposure time to avoid distortion on image size, provoiding identical spatial resolution of the vertical and horizontal axes. The spectral range (λ) recorded was 897–1752 nm at an increment of 3.34 nm, producing hyperspectral images at 256 wavelength channels. The complete image acquisition process was controlled by the SpectralCube software (Spectral Imaging Ltd., Finland). The region of interest (ROI) was selected with MATLAB R2011b (The Mathworks, 2009) and all steps described for spectral analysis were carried out in multivariate analysis software (Unscrambler version 9.7, CAMO, 2007).

3. RESULTS AND DISCUSSION

3.1. Procedure development and optimization

The seeds of each sample were presented to the system spread on the translation stage and conveyed to the FOV of the camera to be scanned line by line. Upon entering the FOV, a hyperspectral image of the sample was acquired and the image in raw format was stored in the computer for further processing.

By examining the acquired hyperspectral images, it was observed that the first five and the last eleven bands of the image had a high level of noise, thus being not useful for spectral data extraction. Therefore, images were cropped to the spectral range of 914 nm to 1715 nm with a total of 240 bands.

A 'white reference' image (*W*, 100% reflectance) was acquired from a white reference ceramic tile, and a 'dark reference' image (*B*, 0% reflectance) was obtained with the light source off and the camera lens completely covered with its opaque cap. The white and dark 'reference' hyperspectral images were used to correct the raw images (R_0) to obtain a relative reflectance image (R) according to the following equation:

$$R = \frac{R_0 - B}{W - B} \qquad (3)$$

Once the images were cropped and corrected as described above, spectra belonging to background and seeds were overlaid. It was selected one band with low spectral reflectance difference between background and seeds and another band with maximum spectral reflectance difference between background and seeds. These bands were 65 and 160 (1127.9 and 1446.4 nm, respectively) (Figure 2). By subtracting these two images it is possible to obtain a new image where this difference appears in grayscale. By applying a thresholding procedure, pixels with the difference being higher than 0.2 were assigned with a value of one to binary mask (Figure 3). Moreover, the shape of the

seeds was eroded with a matrix 1×1 in order to avoid considering pixels in the edges with low-intensity spectrum (González Marcos et al. 2006). For each sample, the region considered by the binary mask was used for the calculation of the mean spectrum as the average of spectra of all pixels within this region.

An average spectrum was extracted from each sample by using the segmentation criterion. Thus, a total of 56 mean reflectance spectra were extracted from grape seeds samples from different varieties. Figure 4 shows the mean and standard deviation spectrum for each class of grape seeds. It can be seen that different categories have different reflectance intensities along some wavelength regions, although with the same pattern. Usually the high reflectance around 950 nm and 1700 nm can be attributed to the high water content of biological materials (Osborne et al. 1993; Murray and Williams, 1987). For each sample, the average spectrum obtained from pixels of HSI in seeds was used for further statistical treatment.

Each hyperspectral image is composed of a large number of contiguous spectral bands. Hence, similar to single-point spectroscopy, a complete reflectance spectrum can be obtained for each pixel in the image. The advantage of hyperspectral imaging lays in the fact that the ROI can be interactively selected during image processing. Statistical approaches are required to extract useful information entrenched in the NIR spectrum and reduce the large number of correlated predictors to a new subset that can explain the maximum variance. According to Wold et al. (1996), selecting optimum wavelengths that carry most of the information may be equally or more efficient than using full spectra. The reduced number of wavelengths is enough to characterize most of the classification tasks (Vila et al. 2005).

In this study, PCA loadings were used for identification of optimal wavelengths that have high influence in each PC (Lawrence et al. 2004). Loadings resulting from PCA of

the spectral data of all samples represent the regression coefficients for each wavelength at the respective principal component and indicate the most dominant wavelengths. New general discriminant analysis (GDA) was carried out using only a few selected wavelengths, and the results were compared with the classification obtained by using the whole spectra.

3.2. Prediction

PLSR was applied to the raw spectral datasets (240 bands) with full cross-validation (leave-one-out) for the original raw spectral data and for pre-processed spectra to predict the stage of maturation. Prediction results using raw spectra were compared with those resulting from the spectral dataset after treatment with different pre-processing methods (SNV, MSC, 1^{st} derivative and 2^{nd} derivative). In addition, 32 samples were randomly selected as a training set (calibration set) and 22 samples were used as a validation set (prediction set).

The results obtained for raw data and each pre-processing treatment selected are summarized in Table 1. In all cases, the optimal number of LV for establishing the calibration model was five. The coefficient of determination in calibration and cross-validation was high and almost the same in all cases. Moreover, there was no difference between predictions in raw spectra and predictions in pre-processed spectra. It denotes that there were no important inconsistencies in instrument response as well as in physical characteristics in samples.

The different pre-processing methods facilitated to investigate the performance of the PLSR models. The difference between R^2_C and R^2_{CV} was small and could indicate the lack of overfitting in the model. The coefficient of prediction values obtained for the full cross-validation and testset validation models were comparable in terms of

prediction performance. Results presented in Table 1 show that regression models based on pre-processed spectra performed similarly to the models for the raw spectra. Derivatives transformation provided equivalent results compared to other pre-processing methods, with comparable R^2_{CV} and RMSECV for the prediction models.

Nevertheless, none of the aforementioned pre-treatment methods inferred in significant improvement of the predictive ability in comparison to the raw spectral data, thus the negative impact of the scatter effects on the regression quality was not confirmed. Therefore, it can be assumed from these results that the applied pre-treatments were not effective for this particular dataset. Given that the predictive ability of the models was similar to that obtained with the original data; it is therefore feasible to assume that the raw spectra extracted using the proposed segmentation approach can be used as representative sample for further analyses. The suitability of PLSR for predicting properties from NIR hyperspectral images of foodstuffs agrees with others similar studies (Barbin et al. 2012; Menesatti et al. 2009).

The loadings of the first three principal components were used for wavelength selection because these three principal components were responsible for 99% of the variance in the spectral data. The wavelengths corresponding to higher module values (peaks and valleys) at these particular principal components were selected as candidates for optimum wavelengths (Figure 5). Six optimum wavelengths (928, 940, 1148, 1325, 1620 and 1656 nm) were thus identified for discrimination purposes.

PC1 explains 95.99% of the total variance in the samples but only one wavelength (928 nm) was selected from this component, while three wavelengths were selected from PC2 (1148, 1325 and 1620 nm), and two from PC 3 (940 and 1656 nm).

Due to the fact that chemical bonds absorb light energy at specific wavelengths, some compositional information can be determined from the reflectance spectra. The spectral

information is repeated through the successive overtones and combination regions. Five of the six optimum wavelengths selected by PCA loadings are due to C–H stretching third (928 and 940 nm), second (1148 nm) and first overtones (1620 and 1652 nm), respectively (Osborne et al. 1993). However, the NIR spectrum contains information from all the chemical constituents of the sample and direct interpretation of the spectral reflectance values is difficult for complex materials such as intact crop seeds (Williams and Norris, 2001).

Figure 6 shows the scores for PC1 and PC2. In this graph, it is not easy to distinguish among varieties. Notwithstanding, a trend in time can be observed along *x* axis (PC1). In this sense, the stage of maturation could be explained by the first principal component. In the plot of PC1 vs time (Figure 7), the dependency can be expressed by its slope. A high slope (in absolute values) means that PC1 is influenced by the time in a more extensively manner. This way, the changes in this principal component are greater in the early stages. Except for Zalema (sand), there is a high dependency between PC1 scores and sampling date. In order to distinguish among varieties, the scores for PC2 and PC3 were also plotted. In Figure 8a, seeds from red grapes (Tempranillo and Syrah) have positive scores in PC3 while seeds from white variety (Zalema) have negative scores. The third principal component also divides the red varieties in two clusters with Tempranillo having higher scores than Syrah. Regarding the variety of Zalema, PC2 splits the grape seeds into two clusters according to the kind of soil (sand and clay). As shown in Figure 8b, the scores from PCA using the six selected wavelengths divides the red and white varieties similarly as using the whole spectrum. However, the division between the two red varieties is not as clear as by using full spectra. A clearer classification can be observed in Figure 9. It shows a three-dimensional scatterplot containing the three first principal components. The principal component analysis

successfully separated the four types of samples used in this study. Although there was a lack of accuracy in some samples, this hyperspectral technique could distinguish two treatments such as kind of soil as it had been previously demonstrated (Karimi et al. 2012).

Since hyperspectral imaging contains the spatial distribution of reflectance, the principal component analysis was also applied in images instead of the average spectra in each sample. In this case, scores appear as intensity in grey-scale in each image. The first principal components have the same meaning as that by considering the average spectrum (Figure 10). PC1 describes evolution in time (seeds appear clearly brighter in the first stages and become darker in the last ones). Once more, this evolution was very weak for this component in Zalema (sand). PC2 also weakly described evolution in time. However, PC2 together with PC3 is useful for classification among varieties.

To verify the potential of selected optimum wavelengths for grape seeds discrimination, GDA was conducted on the reflectance spectral data using the full spectral range (240 wavelengths) and only the optimum wavelengths selected (6 wavelengths).

Table 2 and Table 3 show the classification matrix for the GDA of grapes among varieties. Using the full spectra was possible to classify 100% of the samples according to their variety. The result using only six selected wavelengths is lower, although still satisfactory, since it provided an accuracy of more than 96%. It is clear that optimum wavelengths have a great discrimination power for distinguishing among grape seeds varieties. Acquiring images at those particular wavelengths would reduce image acquisition and processing time and could be useful for establishing a multispectral system for further studies. Due to the similar appearance in red grapes, the unique spectral profile of each variety could be a practical tool to assess the authenticity among

varieties (Woodcock et al. 2008). This could be useful in routine inspections by Designations of Origin to wineries.

4. CONCLUSIONS

The methodology of acquiring hyperspectral images in grape seeds was established. Likewise, the sample presentation as intact seeds and the segmentation criterion chosen provided a suitable way for extracting the mean spectrum of each set of seeds in one sample. The PLSR model applied was able to predict the stage of maturation of a sample based on spectral features as the predictor variables with the coefficients of determination being higher than 0.95. Moreover, the way to acquire the images eliminated the need for pre-processing of images for correction on scattered pixels.

Both PCA and GDA methods were able to characterize the grape seeds according to their varieties. Within the same variety (Zalema), these methods could distinguish between the two kinds of soil where vines were cultivated. The bands containing the most relevant chemical information according to the literature agreed with the bands selected by loadings in PCA.

ACKNOWLEDGEMENTS

This work was supported by the projects AGL2011-30254-C02 (Ministerio de Economía y Competitividad, Gobierno de España), P10-AGR6331 (Consejería de Economía, Innovación, Ciencia y Empresa, Junta de Andalucía), and the concession of the fellowship (BES-2009-025429).

388 **REFERENCES**

389

390 Barbin, D.F., ElMasry, G., Sun, D.W., and Allen, P., 2012. Predicting quality and
391 sensory attributes of pork using near-infrared hyperspectral imaging. Anal Chim Acta.
392 719, 30-42.

393 Barnes, R.J., Dhanoa, M.S., and Lister, S.J., 1989. Standard Normal Variate
394 Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Appl
395 Spectrosc. 43, 772-777.

396 Baye, T.M., Pearson, T.C., and Settles, A.M., 2006. Development of a calibration to
397 predict maize seed composition using single kernel near infrared spectroscopy. J Cereal
398 Sci. 43, 236-243.

399 Burger, J. and Geladi, P., 2006. Hyperspectral NIR imaging for calibration and
400 prediction: a comparison between image and spectrometer data for studying organic and
401 biological samples. Analyst 131, 1152-1160.

402 CAMO, 2007. Unscrambler. CAMO, Trondheim, Norway

403 Choudhary, R., Mahesh, S., Paliwal, J., and Jayas, D.S., 2009. Identification of wheat
404 classes using wavelet features from near infrared hyperspectral images of bulk samples.
405 Biosystems Eng. 102, 115-127.

406 ElMasry, G., Sun, D.W., and Allen, P., 2011. Non-destructive determination of water-
407 holding capacity in fresh beef by using NIR hyperspectral imaging. Food Res Int. 44,
408 2624-2633.

409 Engelbrecht, P., Manley, M., Williams, P.J., Toit, G.D., and Geladi, P., 2010. Pre-
410 germination detected in whole cereal grains using near infrared hyperspectral imaging.

411 Proceedings of the CST SA - ICC International Grains Symposium Quality and Safety
412 of Grain Crops and Foods, pp. 123-127.

413 Escribano-Bailón, T., Alvarez-Garcia, M., Rivas-Gonzalo, J.C., Heredia, F.J., and
414 Santos-Buelga, C., 2001. Color and Stability of Pigments Derived from the
415 Acetaldehyde-Mediated Condensation between Malvidin 3-O-Glucoside and (+)-
416 Catechin. J Agr Food Chem. 49, 1213-1217.

417 Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn Lett. 27, 861-874.

418 Ferrer-Gallego, R., Hernández-Hierro, J.M., Rivas-Gonzalo, J.C., and Escribano-Bailón,
419 M.T., 2010. Feasibility study on the use of near infrared spectroscopy to determine
420 flavanols in grape seeds. Talanta 82, 1778-1783.

421 Gawel, R., 1998. Red wine astringency: a review. Aust J Grape Wine Res. 4, 74-95.

422 Geladi, P., MacDougall, D., and Martens, H., 1985. Linearization and Scatter-
423 Correction for Near-Infrared Reflectance Spectra of Meat. Appl Spectrosc. 39, 491-500.

424 González Marcos, A., Martínez de Pisón Ascacíbar, F.J., Pernía Espinoza, A.V., Alba
425 Elías, F., Castejón Limas, M., Ordieres Meré, J., and Vergara González, E., 2006.
426 Segmentación, in: Universidad de la Rioja. Servicio de Publicaciones (Eds.) Técnicas y
427 algoritmos básicos de visión artificial, Logroño, pp 55-72.

428 González-Miret, M.L., Escudero-Gilete, M.L., and Heredia, F.J., 2006. The
429 establishment of critical control points at the washing and air chilling stages in poultry
430 meat production using multivariate statistics. Food Control 17, 935-941.

431 Isaksson, T. and Næs, T., 1988. The Effect of Multiplicative Scatter Correction (MSC)
432 and Linearity Improvement in NIR Spectroscopy. Appl Spectrosc. 42, 1273-1284.

433 Johnson, D.E., 2000. Métodos multivariados aplicados al anáílisis de datos.
434 International Thomson Editores, S. A., Madrid.

435 Kaihara, M., Takahashi, T., Akazawa, T., Sato, T., and Takahashi, S., 2002. Application
436 of near infrared spectroscopy to rapid analysis of coals. Spectrosc Lett. 35, 369-376.

437 Karimi, Y., Maftoonazad, N., Ramaswamy, H., Prasher, S., and Marcotte, M., 2012.
438 Application of Hyperspectral Technique for Color Classification Avocados Subjected to
439 Different Treatments. Food Bioprocess Tech. 5, 252-264.

440 Kjeldahl, K. and Bro, R., 2010. Some common misunderstandings in chemometrics. J
441 Chemometr. 24, 558-564.

442 Lawrence, K.C., Windham, W.R., Park, B., Smith, D.P., and Poole, G.H., 2004.
443 Comparison between visible/NIR spectroscopy and hyperspectral imaging for detecting
444 surface contaminants on poultry carcasses. in: Bennedsen, B.S., Chen, Y.R., Meyer,
445 G.E., Senecal, A.G., and Tu, S.I. (Eds.), Monitoring Food Safety, Agriculture, and Plant
446 Health. Proceeding of SPIE, Providence, RI, USA, pp. 35-42.

447 Menesatti, P., Zanella, A., D'Andrea, S., Costa, C., Paglia, G., and Pallottino, F., 2009.
448 Supervised Multivariate Analysis of Hyper-spectral NIR Images to Evaluate the Starch
449 Index of Apples. Food Bioprocess Tech. 2, 308-314.

450 Munck, L. and Møller, B., 2004. A new germinative classification model of barley for
451 prediction of malt quality amplified by a near infrared transmission spectroscopy
452 calibration for vigour "on line" both implemented by multivariate data analysis. J I
453 Brewing. 110, 3-17.

454 Murray, I. and Williams, P.C., 1987. Chemical principles of near-infrared technology.
455 in: Williams, P. and Norris, K. (Eds.), Near-Infrared Technology in the Agricultural and
456 Food Industries. American Society of Cereal Chemists, St. Paul, Minnesota, pp. 17-34.

457 Nicolaï, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., and
458 Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by
459 means of NIR spectroscopy: A review. Postharvest Biol Tec. 46, 99-118.

460 Osborne, B.G., Fearn, T., and Hindle, P.T., 1993. Practical NIR Spectroscopy With
461 Applications in Food and Beverage Analysis 2 ed. Longman Group, United Kingdom.

462 Park, B., Chen, Y.R., Hruschka, W.R., Shackeford, S.D., and Koohmaraie, M., 2001.
463 Principal component regression of near-infrared reflectance spectra for beef tenderness
464 prediction. T Am Soc Agr Eng. 43, 609-615.

465 Pizarro, C., Esteban-Díez, I., Nistal, A.J., and González-Sáiz, J.-M., 2004. Influence of
466 data pre-processing on the quantitative determination of the ash content and lipids in
467 roasted coffee by near infrared spectroscopy. Anal Chim Acta. 509, 217-227.

468 Rousseau, J. and Delteil, D., 2000. Présentation d'une méthode d'analyse sensorielle des
469 raisins. Principe, méthode et grille d'interprétation. Rev Fr OEnol. 183, 10-13.

470 Shahin, M.A. and Symons, S.J., 2011. Detection of Fusarium damaged kernels in
471 Canada Western Red Spring wheat using visible/near-infrared hyperspectral imaging
472 and principal component analysis. Comput Electron Agr. 75, 107-112.

473 The Mathworks, 2009. MATLAB. The MathWorks Inc., Natik, Massachusetts

474 Thorngate, J.H. and Singleton, V.L., 1994. Localization of Procyanidins on Grape
475 Seeds. Am J Enol Viticult. 45, 259-262.

476 Vila, J., Calpe, J., Pla, F., Gómez, L., Connell, J., Marchant, J., Calleja, J., Mulqueen,
477 M., Muñoz, J., and Klaren, A., 2005. SmartSpectra: Applying multispectral imaging to
478 industrial environments. Real-Time Imaging. 11, 85-98.

479 Williams, P., Geladi, P., Fox, G., and Manley, M., 2009. Maize kernel hardness
480 classification by near infrared (NIR) hyperspectral imaging and multivariate data
481 analysis. Anal Chim Acta. 653, 121-130.

482 Williams, P. and Norris, K., 2001. Near-Infrared Technology: In the Agricultural and
483 Food Industries 2 ed. American Association of Cereal Chemists, St Paul, MN.

484 Windig, W., Shaver, J., and Bro, R., 2008. Loopy MSC: A Simple Way to Improve
485 Multiplicative Scatter Correction. Appl Spectrosc. 62, 1153-1159.

486 Wold, J.P., Jakobsen, T., and Krane, L., 1996. Atlantic Salmon Average Fat Content
487 Estimated by Near-Infrared Transmittance Spectroscopy. J Food Sci. 61, 74-77.

488 Woodcock, T., Fagan, C., O'Donnell, C.P., and Downey, G., 2008. Application of Near
489 and Mid-Infrared Spectroscopy to Determine Cheese Quality and Authenticity. Food
490 Bioprocess Tech. 1, 117-129.

491 Xing, J., Symons, S., Shahin, M., and Hatcher, D., 2010. Detection of sprout damage in
492 Canada Western Red Spring wheat with multiple wavebands using visible/near-infrared
493 hyperspectral imaging. Biosystems Eng. 106, 188-194.

494
495

496

FIGURE CAPTIONS

Figure 1. Schematic diagram of the hyperspectral imaging system.

Figure 2. Spectra of some pixels from background and seeds and selected band to create the segmentation mask.

Figure 3. Segmentation steps: (a) image of relative reflectance at 1127.9, (b) and 1446.4 nm, (c) image resulting from the subtraction of the two images, and (d) the resulting binary mask after thresholding segmentation and erosion.

Figure 4. Average spectra for the respective samples of each variety. Shaded area represents the standard deviation in each wavelength.

Figure 5. Loadings of the first three principal components showing the selected wavelengths.

Figure 6. Scatterplot of scores for PC1 and PC3. The direction of the arrows indicates the trend in time in each variety.

Figure 7. Dependency of PC1 with date.

Figure 8. Scatterplots of scores for PC2 and PC3 (a) using full spectra (240 bands) and (b) using the selected wavelengths (6 bands).

Figure 9. Three-dimensional scatterplots of scores for PC1, PC2, and PC3 (a) using full spectra and (b) using the selected wavelengths (6 bands).

Figure 10. First three score images obtained from PCA for the hyperspectral images of grape seeds. Horizontal scale shows evolution along maturation.