Talanta 103 (2013) 276–284

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Classification of oat and groat kernels using NIR hyperspectral imaging

Silvia Serranti [a,*], Daniela Cesare [a], Federico Marini [b], Giuseppe Bonifazi [a]

[a] Department of Chemical Engineering Materials & Environment Sapienza—Università di Roma Via Eudossiana 18, 00184 Rome, Italy
[b] Department of Chemistry Sapienza—Università di Roma P.le Aldo Moro 5, 00185 Rome, Italy

ARTICLE INFO

Article history:
Received 26 July 2012
Received in revised form
9 October 2012
Accepted 13 October 2012
Available online 27 October 2012

Keywords:
Near infrared (NIR) hyperspectral imaging
Oats
Principal component analysis (PCA)
Partial least squares-discriminant analysis (PLS-DA)
Sorting
Quality control

ABSTRACT

An innovative procedure to classify oat and groat kernels based on coupling hyperspectral imaging (HSI) in the near infrared (NIR) range (1006–1650 nm) and chemometrics was designed, developed and validated. According to market requirements, the amount of groat, that is the hull-less oat kernels, is one of the most important quality characteristics of oats. Hyperspectral images of oat and groat samples have been acquired by using a NIR spectral camera (Specim, Finland) and the resulting data hypercubes were analyzed applying Principal Component Analysis (PCA) for exploratory purposes and Partial Least Squares-Discriminant Analysis (PLS-DA) to build the classification models to discriminate the two kernel typologies. Results showed that it is possible to accurately recognize oat and groat single kernels by HSI (prediction accuracy was almost 100%). The study demonstrated also that good classification results could be obtained using only three wavelengths (1132, 1195 and 1608 nm), selected by means of a bootstrap-VIP procedure, allowing to speed up the classification processing for industrial applications. The developed objective and non-destructive method based on HSI can be utilized for quality control purposes and/or for the definition of innovative sorting logics of oat grains.
© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Oats is mainly utilized for animal feeding, human consumption and industrial uses. Concerning human consumption, its high nutritional value combined with its characteristic flavor has contributed to secure an important place in the breakfast cereals group and other processed food.

The oat kernel consists of a caryopsis, or groat, and a hull, consisting of leaf-like structures that tightly enclose the groat and provide protection during seed growth [1]. The hull, that contributes to about 30% of the total kernel weight, is usually removed before the grain can be processed for human consumption. Groat percentage, that is the amount in weight of hull-less kernels on the total oats, is one of the most important quality characteristics of oats.

There are three basic mechanical methods for dehulling oats, including the impact dehuller, the compressed-air dehuller, and the wringer dehuller [2].

Commercial oat processing generally utilizes the impact dehuller for this purpose. During impact dehulling, grains are fed into the top of a spinning rotor, which expels the grain against the wall of the dehuller. The force of the impact causes the hull to break away from the groat and the lighter hulls are then removed by aspiration.

Dehulling efficiency, that is the proportion of oat grains that are dehulled during a single pass through an impact dehuller, increases with increasing impact rotor speed, but the increased rotor speed also produces more broken groats. High proportions of broken groats can reduce the value of the final products, especially in those food applications requiring intact groats for improved value. To maximize unbroken groat yield, a suitable rotor speed must be used for maximal dehulling efficiency but minimal groat breakage. Several oats characteristics affect the dehulling efficiency, such as kernel density, oil and protein concentration in groat, grain moisture, etc.

Due to the previous mentioned reasons, even after dehulling, groat is not completely clean and variable amount of oats are still present in the products. Common optical sorting methods are not so efficient in refining this process, due to the similar reflectance properties in the visible region of oats and groat.

NIR spectroscopy is a widespread method to evaluate chemical characteristics of cereals, especially for quality screening [3]. In particular, evaluation of groat percentage in oats was studied by conventional NIR spectroscopy, providing good prediction results [4].

The use of hyperspectral imaging in the near infrared field (1000–1700 nm) has been investigated in this study, in order to develop a new method for classification of oats and groat kernels both for quality control of the oat products and for implementation

* Corresponding author. Tel.: +39 6 44585360; fax: +39 6 44585618.
E-mail address: silvia.serranti@uniroma1.it (S. Serranti).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.10.044

of innovative on-line sorting strategies. The study was carried out to reach two different goals: (i) to verify the possibility of recognition of oat and groat by NIR-HSI and (ii) once verified the first condition, to select important wavelengths for further development of an on-line inspection system.

2. Hyperspectral imaging technique

HSI is based on the utilization of an integrated hardware and software architecture able to digitally capture and handle spectra, as an image sequence, as they result along a pre-defined alignment on a surface sample properly energized [5,6]. The spatial and spectral information, obtained at the same time from the investigated object, are arranged in a "hypercube", a 3D dataset characterized by two spatial dimensions and one spectral dimension. Considering that in a hyperspectral image the spectrum of each pixel can be analyzed, HSI is the non-destructive technology providing the most accurate and detailed information extraction. According to the different wavelength of the source and the spectral sensitivity of the device, several physical–chemical characteristics of a sample can be investigated and analyzed. For these reasons, HSI techniques represent an attractive solution for characterization, classification and quality control of different materials in several industrial sectors. In these last years HSI has rapidly emerged and fast-grown especially in food inspection [7,8], in the detection of contaminant in agrofood products (for instance meat and bone meals in feed, in relation to the BSE crisis) [9,10], in the pharmaceutical sector [11,12], in medicine [13,14], in artworks [15] and in polymer science [16]. Studies have been also carried out in solid waste sectors, e.g., with reference to glass recycling [17], characterization of car-fluff [18] and bottom ash from municipal solid waste incinerators [19], compost products quality control [20] and post-consumer polyolefins classification [21,22]. In the framework of the quality control of cereals, especially of maize and wheat, different applications of HSI involving, for instance, the assessment of fungal contamination [23,24], the detection of insect-induced damages [25], or the determination of the hardness of kernels [26], have been reported in the literature.

The studies carried out on the application of HSI techniques for material classification and inspection are increasing every day, demonstrating that this technique is a very smart and promising analytical tool for quality control. Despite these advantages, HSI is still difficult to be systematically applied, especially in real-time industrial applications, because of the huge amount of data contained in spectral images, requiring too long computation time in comparison to the high-speed of processing required when dealing for example with particles moving on a conveyor belt to be recognized and/or selected. One way of overcoming this problem is the identification of a few most useful wavelengths, reducing in this way the acquisition and the data analysis time [27].

3. Materials and methods

3.1. Oat samples

In this study, three different sets of oat samples have been selected and utilized, the first one composed of whole oat kernels, the second one of only groat (de-hulled) and the third one constituted by oat and groat mixed together.

The three sets are representative of the products involved in a separation process for the refining of the de-hulling operation performed on oats. The oat and groat mix is the input of the process, whereas the two separated types are the outputs.

The spectra of the three samples have been acquired by hyperspectral imaging according to different experimental set-up, that is:

- Experimental set-up n.1. The two separated oat and groat samples have been acquired in bulk. About 10 g of each sample have been selected, corresponding to around 100 single grains of oat and 200 of groat and placed in two Petri dishes of 5 cm diameter. These sample sets have utilized to build the classification model;
- Experimental set-up n.2. Three different and separated parallel lines, constituted by 10 to 15 single seeds each, of known typology, have been acquired and processed. The first line was constituted of oat kernels, the second of groat kernels and the third of mixed kernels. This second acquisition was carried out in order to validate the classification model developed through the acquisition and processing performed in the previous experimental set-up.
- Experimental set-up n.3. The mixed oat and groat sample, representative of the feed of the refining process, was acquired in order to test the developed model on a real bulk sample.

3.2. The hyperspectral imaging (HSI) platform

The HSI acquisitions of oat samples have been carried out at the Laboratory for Particles and Particulate Solids Characterization (Latina, Italy) of the Department of Chemical Engineering, Materials and Environment ("Sapienza" University of Rome).

A specifically designed hyperspectral imaging based platform (DV srl, Italy) was utilized to perform all the analyses. The HSI based detection architecture was realized to allow not only static, but also dynamic analysis, that is the possibility to carry out tests on particle flow streams transported on a conveyor belt in order to perform, at laboratory scale, particles *on-line* detection in a sorting and/or quality control perspective.

The platform, in terms of hardware components, is based on a controlled conveyor belt (width = 26 cm and length = 160 cm) with adjustable speed (variable between 0 and 50 mm/s). The utilized acquisition system is a NIR Spectral Camera™ (Specim, Finland), embedding an ImSpector N17E™ imaging spectrograph working in spectral range from 1000 to 1700 nm, with a spectral sampling/pixel of 2.6 nm, coupled with a Te-cooled InGaAs photodiode array sensor (320 × 240 pixels) with pixel resolution of 12 bits. A diffused light cylinder source, providing the required energy for the sensing unit, was set-up. The cylinder, aluminum internally coated, embeds five halogen lamps producing a continuous spectrum signal optimized for spectra acquisition in the NIR wavelength range. The device works as a push-broom type line scan camera allowing the acquisition of spectral information for each pixel in the line [5]. The transmission diffraction grating and optics provide high-light throughput and high quality and distortion-less image for the device. The result of acquisition is a digital image where each column represents the discrete spectrum values of the corresponding element of the sensitive linear array.

The device is controlled by a PC unit equipped with the Spectral Scanner™ v.2.3 acquisition/pre-processing software, specifically developed to handle the different units and the sensing device constituting the platform and to perform the acquisition and the collection of spectra. The software was designed as a flexible architecture to be easily integrated with new software modules embedding new characterization and/or classification tools.

3.3. Hyperspectral image acquisition

Hyperspectral images of the oat kernels have been acquired in the 880–1720 nm wavelength range, with a spectral resolution of 7 nm, for a total of 121 wavelengths. The spectrometer was coupled to a 15 mm lens. The images were acquired scanning the investigated sample line by line. The image width was 320 pixels, while the number of frames was variable from 200 to 350, depending on the length of the sample.

Calibration was performed recording two images for black and white references. The black image (B) was acquired to remove the effect of dark current of the camera sensor, turning off the light source and covering the camera lens with its cap. The white reference image (W) was acquired for a standard white ceramic tile under the same conditions of the raw image. Image correction was thus performed adopting the following equation (I):

$$I = \frac{I_0 - B}{W - B} \times 100 \quad (1)$$

where I is the corrected hyperspectral image in a unit of relative reflectance (%), I_0 is the original hyperspectral image, B is the black reference image (~0% reflectance) and W is the white reference image (~99.9% reflectance). All the corrected images were then used to perform the HSI based analysis, that is to extract spectral information, to select the effective wavelengths and for the final classification purposes.

4. Spectral data analysis

Hyperspectral data were analyzed using the PLS_Toolbox (Version 6.5.1, Eigenvector Research, Inc., Wenatchee, WA) under Matlab® environment (Version 7.11.1, The Mathworks, Inc., Natick, MA). Data were processed by standard chemometric methods [28,6].

4.1. Spectral preprocessing

First, the number of spectral variables was reduced from 121 to 93, the corresponding wavelength interval being cut to 1006–1650 nm, in order to eliminate unwanted effects due to background noise at the beginning and at the end of the frequency range. Successively, prior to any exploratory or classification analysis, spectral preprocessing was carried out. Indeed, different unwanted contributions (background or other signals which are interferents with respect to the multivariate models to be built) can contribute significantly to the variability observed in spectral signals. In this respect, the use of spectral preprocessing techniques helps reducing the impact of these sources of variability on the overall signal, allowing a more accurate modeling of the investigated phenomena and a better and clearer interpretation of the results. In this study, different algorithms for spectral pretreatment (such as SNV [29], detrending [29], first and second derivative, according to the Savitzky–Golay method [30], and their combinations) were initially tested, but the best results were obtained through the use of the "generalized least squares weighting" (GLSW) algorithm [31], that calculates a filter matrix based on the differences between pairs or groups of samples, which should otherwise be similar. In the case of problems like the one in the present research, where samples from different categories (groat and oat) are present, similar samples would be the members of a given class. Therefore, the goal of GLSW is to remove the within-class variance as much as possible without reducing at the same time the between-class one. More in detail, the algorithm proceeds by centering the data coming from the different categories to their own class mean and use this class-centered data matrix to compute the filter. If $\mathbf{X}_i$ is the submatrix containing the data from class i, the class-centered data matrix $\mathbf{X}_c$, is computed according to:

$$\mathbf{X}_c = \begin{bmatrix} \mathbf{X}_1 - \mathbf{1}_1 \mathrm{m}_1^T \\ \vdots \\ \mathbf{X}_g - \mathbf{1}_g \mathrm{m}_g^T \end{bmatrix} \quad (2)$$

where $\mathbf{m}_i$ is the row vector containing the mean spectrum for the samples of class i, and $\mathbf{1}_i$ is a column vectors of all one having the same length as the number of samples in the same category. Then the matrix $\mathbf{X}_g$ is used to compute a covariance matrix $\mathbf{C}$, which is decomposed according to an SVD scheme:

$$\mathbf{C} = \mathbf{X}_c^T \mathbf{X}_c = \mathbf{V}\mathbf{S}\mathbf{V}^T \quad (3)$$

where $\mathbf{V}$ and $\mathbf{S}$ are the eigenvectors and the diagonal matrix of singular values, respectively. Lastly, the filter matrix $\mathbf{G}$ is calculated using a weighted, ridged version of the singular value matrix:

$$\mathbf{G} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T \quad (4)$$

where:

$$\mathbf{D} = \sqrt{\frac{\mathbf{S}}{\alpha} + \mathrm{I}_D} \quad (5)$$

$\mathbf{I}_D$ and α being the identity matrix of appropriate dimensionality and the weighting parameter, respectively. The adjustable parameter, α, defines how strongly GLSW downweights interferences: adjusting α towards larger values (typically above 0.001) decreases the effect of the filter, while smaller αs (typically 0.001 and below) apply more filtering.

4.2. Exploratory data analysis by principal component analysis

Hyperspectral cameras produce a huge amount of data (the hypercubes considered in this study, after variable reduction are $320 \times 240 \times 93$) so chemometric processing is needed for data exploration and modeling. The first stage of chemometric processing is usually exploratory data analysis, which is carried out through the help of principal component analysis (PCA) [32]. PCA compresses the data by projecting the samples into a low dimensional subspace, whose axes (the principal components, PCs) point in the directions of maximal variance. As the main sources of data variability are concentrated in a few variables (often 2 or 3), from the observation of the distribution of the samples onto the PC space it is possible to analyze their common features and/or their grouping. On the other hand, inspection of the loadings allows to interpret the observed differences and similarities among the samples in terms of the original spectral fingerprint.

4.3. Partial least square-discriminant analysis (PLS-DA)

Principal component analysis is a powerful method for data exploration, being able to highlight the presence of trends or clusters among samples. However, PCA is an unsupervised technique and it cannot be used for building predictive model, for instance to classify samples in one or another category: in the latter case, a supervised pattern recognition approach should be adopted.

In particular, as the aim of our study was to develop a method for the identification of oat or groat samples through the use of NIR hyperspectral imaging, the second data analytical step involved the construction of chemometric classification models by means of partial least squares-discriminant analysis (PLS-DA) [33]. As all the other discriminant classification techniques,

PLS-DA assigns an unknown sample to one (and only one) of the available categories based on its spectral fingerprint. As the name itself suggests, the core of the PLS-DA approach is the use of partial least squares regression [34], which operates a bilinear decomposition of both the *X*- and *Y*-spaces, under the assumption that a relationship between the two internal spaces exists, to compute the model parameters. Since PLS-DA modeling implies the projection of the data onto a subspace of abstract (or "latent") variables (analogously to what happens with PCA), there is the need of estimating the optimal dimensionality of this subspace. To this purpose, given the high number of observations in our training set, 10-fold cross-validation was used for the selection of the optimal complexity of the classification models. In particular, cross-validation was repeated 20 times after random division of the training samples into the 10 cancellation groups and the overall CV classification error was computed as a function of the number of latent variables: the number of latent variables corresponding to the minimum in the error curve was selected.

Once the model is obtained, it can be applied to whole images for the classification of the individual pixels: the result of PLS-DA applied to hyperspectral images is a "prediction map", where the class of each pixel can be identified using color mapping. The purpose of PLS-DA applied to oat and groat samples was to validate their correct classification using both all the wavelengths and the effective wavelengths selected applying a bootstrap-VIP approach.

4.4. Wavelength selection

Considering that the extracted spectral data from oat kernels images are characterized by a high dimensionality with redundancy among contiguous predictors (wavelengths), in a further stage of this study variable selection was carried out, in order to facilitate and speed up the classification of oat and groat. In particular, to select a minimum number of significant variables, which could be relevant for predicting whether an unknown observation comes from oat or groat, a bootstrap-VIP approach [35] was followed. Variable importance in projection (VIP) [36] is a PLS-based score that expresses whether a predictor is significant in the definition of the *F* latent vectors model for the prediction of a particular response. Mathematically, it is defined according to the formula:

$$\mathrm{VIP}_j = \sqrt{N_{\mathrm{vars}} \frac{\sum_{k=1}^{F} \left(b_k^2 \mathbf{t}_k^T \mathbf{t}_k\right) \left(w_{jk} / \|w_k\|\right)^2}{\sum_{k=1}^{F} \left(b_k^2 \mathbf{t}_k^T \mathbf{t}_k\right)}} \quad (6)$$

where $\mathbf{t}_k$ is the vector of sample scores along the kth latent variable, b_k is the coefficient of the kth PLS inner relationship, N_{var_s} is the number of experimental variables and w_{jk} and $\mathbf{w}_k$ are the weight of the jth variable for the kth LV and the weight vector for the kth LV, respectively. Since the average of squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion to identify the most significant variables.

In this study, bootstrap re-sampling was applied to generate confidence intervals around VIP estimates for each variable. Indeed, bootstrap is a well-known re-sampling method used in statistics to estimate bias and standard errors of parameter estimates: the dataset was randomly re-sampled 500 times with replacement leading to 500 estimates of the parameters of interest (i.e., VIP) after building a PLS-DA model on each of the 500 datasets. The PLS-bootstrap algorithm was used to assess the uncertainty in the VIP metrics, and, accordingly, a wavelength was considered relevant when the lower confidence limit at 95% of its VIP value (obtained from the bootstrap) was above 1.0.

5. Experimental results and discussion

As already pointed out, the main scope of this work is to build a model able to differentiate oat and groat samples using NIR hyperspectral imaging. To this purpose, three images (described in the materials and methods section as Experimental set-up 1–3) have been collected and analyzed. In particular, it has been chosen to use the image corresponding to the experimental set-up 1, as the one to be used for building the models, leaving the other two as external validation sets.

In particular, 2 square ROIs (regions of interest) of about 2×2 cm size, representative of groats and oats, respectively, were selected from the image corresponding to experimental set-up1. The corresponding spectra were arranged in a two dimensional matrix which constituted the training set for this investigation.

5.1. Exploratory data analysis

Starting from the training matrix made up of the spectra from the selected ROIs, exploratory data analysis was performed. In particular, at first the raw signals measured on pixels corresponding to the two different classes in exams, where characterized and compared (Fig. 1a).

The investigated NIR range provides chemical information about cereal composition because most absorption bands observed in this region arise from overtones of C–H, O–H and N–H stretching vibrations [37]. More in details, it can be observed from Fig. 1a that both classes present spectra characterized by the presence of two main bands: the valley at around 1200 nm is related to the second overtone of C–H stretching vibrations while the most intense one at 1450 nm is due to the first overtones of O–H and N–H stretching vibrations and to combination bands of the C–H in the aromatic ring.

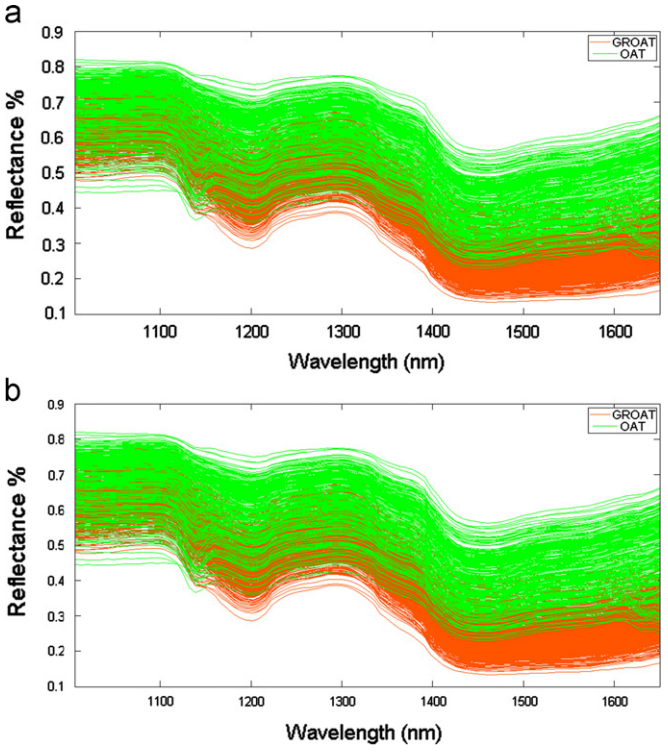

Fig. 1. NIR spectra of whole oat sample (red) and groat sample (green) before (1a) and after (1b) pre-processing by the GLSW algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Then principal component analysis was used for a more thorough characterization of the training data at an exploratory level. As already described in Section 4.1, in order to remove as much as possible the contribution of unwanted sources of variability from the spectral signal, GLS weighting (setting the α parameter in equation 5 to 0.002) was used. The effectiveness of this decluttering filter can be visualized in Fig. 1b, where the raw and the pretreated spectra for the training samples from the two categories oat and groat are reported.

It is apparent from Fig. 1 that the use of GLS weighting for signal preprocessing allows to bring out the differences between oat and groat spectral features, by filtering out the contributions of other sources of variability that otherwise would increase the intra-category variance (as in the case of the raw signals). Moreover, inspection of Fig. 1b shows that the spectral pretreatment also allowed an easier visual identification of the presence of a limited number of outlying pixels (whose profiles are significantly different from all the other spectra of the corresponding class), which could not be pointed out when looking at the raw data. Those pixels, accordingly, were removed from the training set prior to any further analysis.

The training data, pretreated as described, were then used to build a Principal Component model for exploratory purposes. In particular, the projection of the samples onto the first two PCs (accounting for 73.29% of the total variance: 66.32% and 6.97%, respectively) is reported in Fig. 2.

It is evident from Fig. 2 that the two classes are well separated along the first principal component, thus indicating that the spectral fingerprints associated to the pixels carry a discriminant information, which is magnified by the use of GLSW pretreatment. Indeed, the fact that the separation among the categories occurs along the first principal component and that this component accounts for a significant portion of the total variance (more than 65%) confirms that spectral pretreatment was able to effectively filter out a significant part of the signal variability not associated with the class belonging (i.e., not useful for differentiating among oat and groat).

As the first principal component allows to discriminate between the classes, inspection of the loadings for this PC (Fig. 3) can provide an interpretation, by indicating what are the spectral regions which contribute the most to the observed differences. Fig. 3 shows that the highest contribution to the definition of PC1 come from the signals at 1132 and 1433 nm (negative) and those at 1195, 1384 and 1608 nm (positive): since groat pixels have a negative score on PC1, while oat ones fall on the positive side, the former are characterized by higher intensity in the spectral regions corresponding to negative loadings (1132 and 1433 nm) and lower pseudo-absorbance for the other bands (1195, 1384 and 1608 nm). The reverse is true for oat samples.

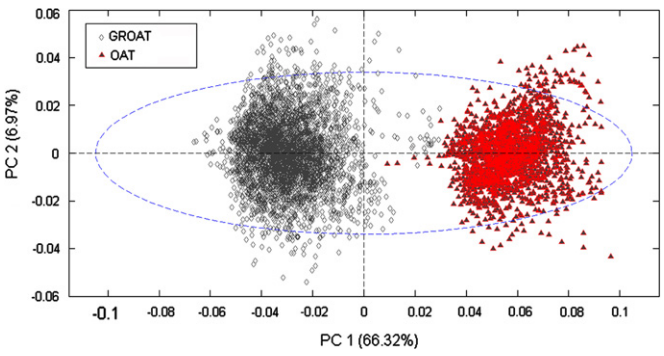

Fig. 2. PCA score plot: projection of the training pixel onto the space spanned by the first two principal components (PC1 vs PC2). Legend: ◇ groat; ▲ oat.

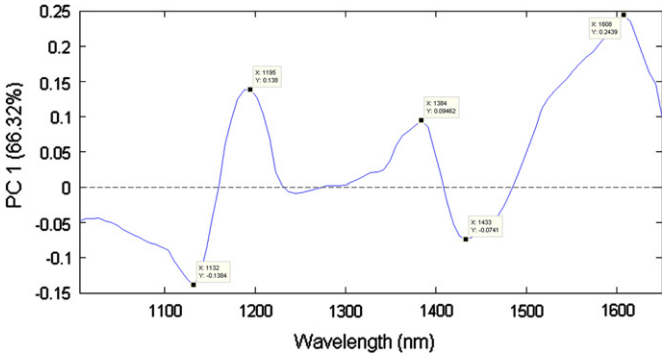

Fig. 3. Loadings of the spectral variables for the definition of the first principal component.

Based on this PC model, exploratory analysis of the whole image (after masking the background) was made by projecting all the pixels onto the PC space defined by the loadings calculated on the training set. When the whole image is projected onto the principal component space, each pixel is described by its scores on the significant PCs. Then, for the sake of exploratory data analysis, two different representations can be adopted. The first one consists in drawing a scatterplot where each pixel is described as a point whose coordinates are its scores onto pairs of principal components, analogously to what already reported in Fig. 2, in the case of the training set only. This representation allows to highlight the presence of clusters, trends or outlying points. In the case of the hypercube corresponding to the experimental set up 1, projection of the whole image onto the PC space computed from the training pixels allowed to identify the absence of outlying pixels and confirmed the observation that the first principal component accounts for the separation between the different classes. Indeed, when projected onto the model computed on the training set, the remaining pixels from the image fell in the cluster corresponding to their respective category (oat or groat).

However, when PCA is used to analyze a hyperspectral image, together with the usual scatterplot of pixels in the score space, another representation of the results is also possible, which allows a better visualization. Indeed, the scores of all points along a single principal component can be refolded back into an image: in these score images, the intensity of the pixels are given by the scores of the pixels themselves onto the particular principal component. In particular, as it was already discussed that separation among the pixels corresponding to the two categories to discriminate occurs along the first principal component, the corresponding score image was further investigated. The score image built from the projection of the hypercube corresponding to Experimental set-up 1 onto the first principal component (after masking the background) is reported in Fig. 4.

With the only exception of some boundary effects, linked to background removal, it can be seen from Fig. 4 how the pixel coloring reflects perfectly the differences between the two classes, as already evidenced by the clear separation in Fig. 2. Indeed, the coloring shows that almost all the pixels corresponding to groat have a negative score value along PC1 (not only the training points but also the other ones), while those corresponding to oat display scores whose values are consistently greater than zero.

5.2. *Classification using PLS-DA on the full wavelength range*

In a successive stage, as the main aim of the study was to build a model able to discriminate between oat and groat samples

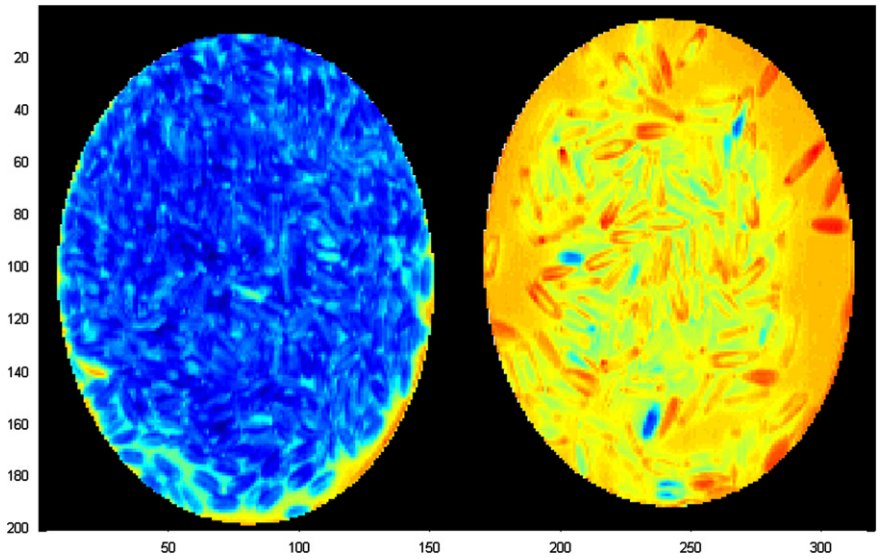

Fig. 4. Score image corresponding to the projection of the hypercube corresponding to the Experimental set-up 1 onto the first principal component PC1. Left: Groat samples; right: oat samples.

based on their hyperspectral fingerprint, classification was performed by means of the PLS-DA algorithm. In particular, as in all images background was masked, a model consisting on two categories only (oat and groat) was built. To this purpose, the set of pixels coming from the image corresponding to Experimental set-up1 and already described in the previous subsection was used as the training set, while the remaining pixels of that image and the other two images were used as external validation sets. In particular, to avoid unbalance between the two categories, 1500 spectra were randomly selected from each of the two ROIs, corresponding to oat and groat, to constitute the final training set. The optimal complexity of the PLS-DA model was chosen based on the minimum classification error in 10-fold cross-validation and it was found to be 7 latent variables. The classification ability of this model was very high, resulting in 99.3% and 100.0% non error rates both in calibration and in cross-validation for groat and oat pixels respectively. When this model was applied to the remaining pixels of the same image (Experimental set-up 1) left out as test set, comparably high predictive ability was obtained: indeed, more than 99% correct classification rate for both categories was obtained in prediction on these validation pixels. Successively, the same PLS-DA model was applied to the other two images left out as further validation sets. In particular, the test image corresponding to the experimental set-up 2 was made of oat and groat grains arranged in three parallel lines of single seeds (Fig. 5a): the first one contains both oat and groat (13 groat and 2 oat kernels), the second one only groat (13 kernels) and the last one only oat (10 kernels). When the PLS-DA model computed on the training pixels was applied to this validation image, again a correct classification rate higher than 99% was obtained for both classes, the only errors in prediction being related to pixels corresponding to the boundaries of the kernels. These results can be visualized in Fig. 5b, in the form of a prediction image, i.e., of an image having the same dimension as the original one but where pixels are colored according to the predicted category. In this case, white pixels correspond to the background, which was masked. It is evident from Fig. 5b that accurate predictions are obtained for both categories and that the few misclassifications that are observed are related to pixels which are at the boundary between the kernels and the masked background.

Finally, the PLS-DA model was applied to the other validation image (corresponding to Experimental set-up 3), which was recorded on a sample constituted by a mix of oat and groat seeds, in order to test the procedure on a "real" bulk sample, as the proposed procedure has been developed with the aim of controlling and assisting the process of de-hulling. Even in this more difficult case, using the PLS-DA model a good prediction of both classes was obtained, as shown in the prediction image of Fig. 6. The green colored pixels (5.4% of the total pixels) correspond to the oat kernels dispersed in the groat ones, that are represented by the red colored pixels (94.6% of the total pixels). As the bulk material was characterized by a kernel distribution of about 94% groat and 6% oat, the recovered pixel percentages, together with the visual inspection of the prediction image, showing that all the misclassification occur only at the boundary of the different kernels, confirm the effectiveness of the proposed method.

5.3. Classification on a reduced set of variables

Having in mind the aim of using the proposed method for on-line process monitoring and/or control, the possibility of building classification models on a reduced number of wavelengths appears promising as it could decrease the time required to acquire and process each spectral image, opening this way interesting scenarios for the design, the development and the set-up of innovative HSI based sorting procedures of oat and groat grains for quality control and/or separation purposes.

Accordingly, in the last stage of our study, this possibility was investigated in depth. In particular, as at the industrial level the sorting machines require the use of few wavelengths, a variable selection procedure to identify three wavelengths to be used for model building was carried out. As described in Section 4.2 a procedure based on the calculation of VIP scores was adopted. Indeed, VIP scores for all the 93 variables were calculated based on the PLS-DA model built on the full wavelength data-set. Successively, confidence intervals around this estimate of the VIPs were computed using a bootstrap procedure with 500 repetitions (Fig. 7). Eventually, the three wavelengths for which the lower confidence limit (at 95%) of the VIP score was highest were selected: 1132, 1195 and 1608 nm.

The three variables selected based on bootstrapped VIPs are consistent with the wavelengths identified as significant by inspection of the PCA loadings (see Section 5.1) and, as already discussed, correspond to chemically meaningful spectral signals.

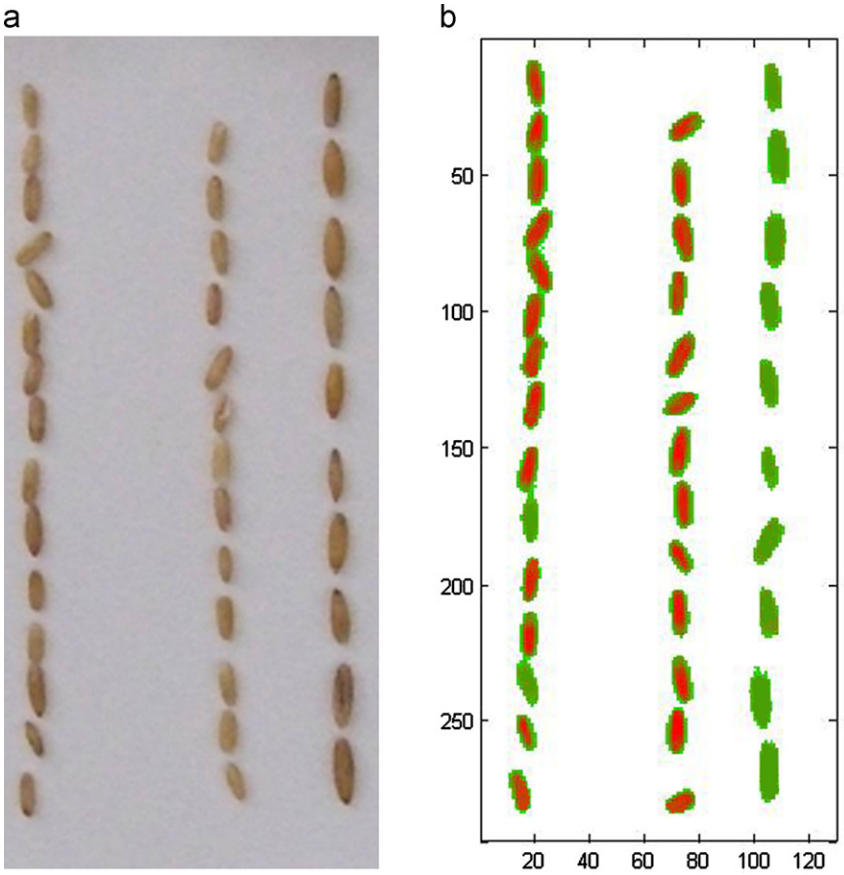

Fig. 5. Source (a) and prediction image (b) based on PLS-DA model built for the classification of oat and groat in lines using 93 wavelengths (red: groat; green: oat). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

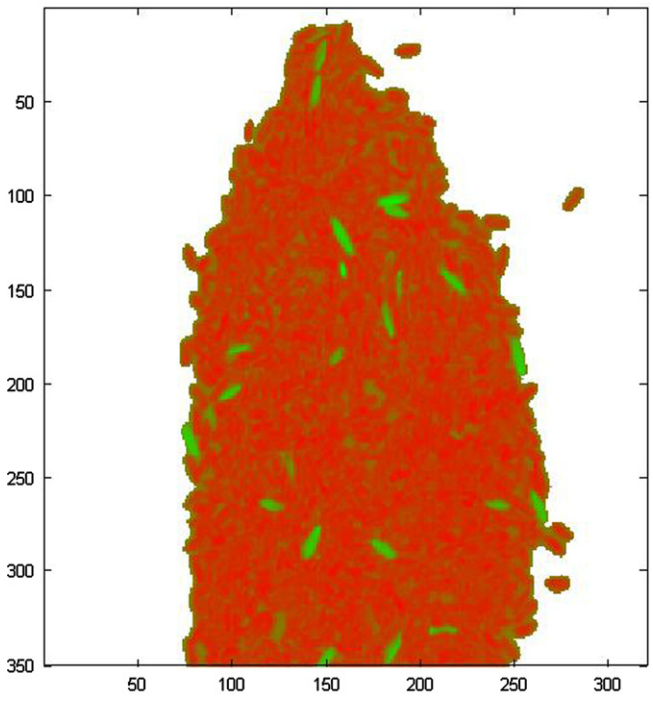

Fig. 6. Prediction image based on PLS-DA model built for the classification of oat and groat in the mix sample representative of the feed of the refining process, obtained using 93 wavelengths (red: groat; green: oat). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

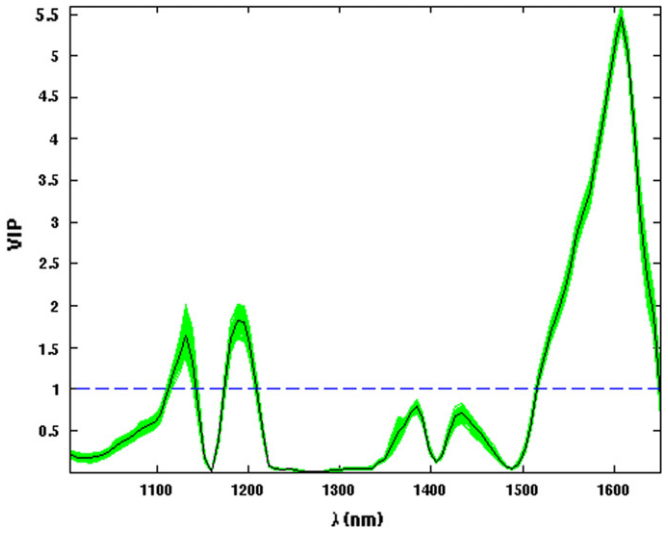

Fig. 7. PLS-DA modeling: VIP scores for the 93 spectral variables (black line) and their confidence intervals estimated by the described bootstrap procedure. The dashed horizontal line indicates the threshold value of 1.

PLS-DA model was then built on the same training pixels used for the full wavelength case but including only the three selected variables. Also in this case, the optimal complexity of the model was estimated as that leading to the minimum error in 10-fold cross-validation and was found to be 1 LV. The classification ability of this model was slightly worse than that obtained using

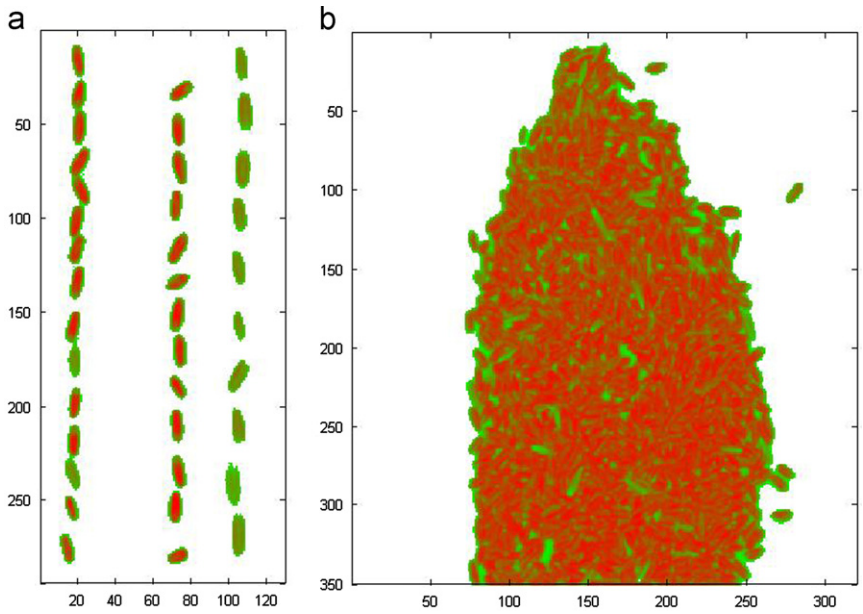

Fig. 8. Prediction images based on PLS-DA model built for the classification of oat and groat in lines (a) and in bulk (b) using 3 wavelengths (red: groat; green: oat). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

all the variables, but still it was very good. Indeed, 97.1% and 99.0% correct classification rate on groat and oat, respectively, were obtained both in calibration and cross-validation. Accordingly, when this model was applied to the validation images, comparably accurate predictions were obtained, non error rate in prediction being in all cases higher than 97%. In particular, the prediction images corresponding to the experimental set-ups 2 and 3 are reported in Fig. 8.

The prediction image in Fig. 8a shows that all the seeds in lines are correctly classified (green colored: oat, red colored: groat) and that, analogously to what already observed in the case of the model built with all 93 variables, the only mis-predictions occur at the boundaries of the kernel. On the other hand, Fig. 8b indicates that the predictions in case of the bulk sample representing the feed of the refining process are slightly worse than with all the wavelengths but still very good. Indeed, compared to the actual composition (94% groat and 6% oat), the green colored pixels, corresponding to the oat kernels are 7.9%, whereas the red pixels, representative of groat ones, are 92.1% of the total pixels, which is in good accordance. Additionally, as also in this case the mis-predictions occur at the boundary of the grains, if classification was made kernel-wise and not pixel-wise, error rate would decrease further. It is apparent that the use of a reduced set of variables does not affect significantly the classification ability of the model, so that the accuracy and the quality of the predictions does not seem to be worsened to a relevant extent.

6. Conclusions

A procedure based on coupling hyperspectral imaging (HSI) in the near infrared region (1006–1650 nm) to chemometric processing was developed to evaluate oat kernels before and after de-hulling. In particular, classification models to discriminate between oat and groat kernels were built using PLS-DA and allowed to obtain a predictive accuracy near to 100% for both the investigated categories. Moreover, it was demonstrated that models resulting in a comparable accuracy can be built using only a reduced set of three wavelengths (1132, 1195 and 1608 nm). This latter outcome is really promising, as it could allow to significantly decrease the time required to handle the hyperspectral data allowing a real time monitoring of the process. The proposed approach can thus profitably utilized for classification purposes and clearly shows the potentiality of NIR hyperspectral imaging to screen samples according to their spectral information. The results showed as the NIR hyperspectral imaging based approach, coupled with multivariate statistical analysis methods, such as PCA and PLS-DA, holds the advantages to be an objective, rapid and non-destructive technique for the recognition of whole oat and groat. The adoption and implementation of such an approach can be utilized for fast quality control purposes and/or to realize a sorting of the products that is not easy to reach with conventional techniques. Actually, at industrial level, oat and groat sorting is performed utilizing 6–7 shaking tables in series that are not even sufficient to reach the best product quality required by the market. The implementation of the proposed innovative optical approach, only based on one selection unit, could greatly improve oat product quality, reducing at the same time the costs of about 1/3 in terms of equipment and of more than a half in terms of energy.

Acknowledgements

Thanks are given to Mr. Antonio and Andrea Uzzo of SEA s.r.l. (Imola, BO, Italy) for having provided the oat samples. A special thanks to Eng. Laura D'Aniello for the acquisition of samples.

References

[1] D.C. Doehlert, D.P. Wiesenborn, M.S. McMuller, J.-B. Ohm, N.R. Riveland, Cereal Chem. 86 (6) (2009) 653–660.
[2] D.C. Doehlert, M.S. McMuller, Cereal Chem. 78 (6) (2001) 675–679.
[3] B.G. Osborne, J. Near Infrared Spectrosc. 14 (2006) 93–101.
[4] R. Redaelli, N. Berardo, J. Near Infrared Spectrosc. 10 (2002) 103–109.
[5] T. Hyvarinen, E. Herrala, A. Dall'Ava, in: Proceedings of SPIE Electronic Imaging, 3302, San Jose, CA, USA (1998).
[6] P. Geladi, H. Grahn, J. Burger, in: H. Grahn, P. Geladi (Eds.), Techniques and Applications of Hyperspectral Image Analysis, John Wiley & Sons, West Sussex, England, 2007, pp. 1–15.
[7] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Trends Food Sci. Technol. 18 (2007) 590–598.
[8] D.-W. Sun (Ed.), Hyperspectral Imaging for Food Quality Analysis and Control, Academic Press/Elsevier Ed., San Diego CA, 2010.

[9] V. Baeten, J.A. Fernandez Pierna, P. Dardenne, in: H. Grahn, P. Geladi (Eds.), Techniques and Applications of Hyperspectral Image Analysis, John Wiley & Sons, West Sussex, England, 2007, pp. 289–311.
[10] F. Pierna, J.A.V. Baeten, P. Dardenne, E.N. Lewis, J. Dubois, J. Burger, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, UK, 2009, pp. 173–196.
[11] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, S.E.J. Bellc, Eur. J. Pharm. Biopharm. 69 (2008) 10–22.
[12] W. Fortunato de Carvalho Rocha, G. Post Sabin, P.H. Março, R.J. Poppi, Chemom. Intell. Lab. Syst. 106 (2011) 198–204.
[13] R. Jolivot, P. Vabres, F. Marzani, Comput. Med. Imaging Graph. 5 (2011) 85–88.
[14] F. Blanco, M. López-Mesas, S. Serranti, G. Bonifazi, J. Havel, M. Valiente, J. Biomed. Opt. 17 (2012) 076027-1-076027-12.
[15] M. Kubik, in: D. Creagh, D. Bradley (Eds.), Physical Techniques in the Study of Art, Archaeology and Cultural Heritage, Elsevier, New York, 2007, pp. 199–259.
[16] R. Gosselin, D. Rodrigue, C. Duchesne, Comput. Chem. Eng. 35 (2011) 296–306.
[17] G. Bonifazi, S. Serranti, Waste Manage. 26 (2006) 627–639.
[18] G. Bonifazi, S. Serranti, Proc. SPIE 6377, U151-U160, Bellingham, WA, USA http://dx.doi.org/DOI:10.1117/12.684661 (2006).
[19] G. Bonifazi, S. Serranti, Proc. SPIE 6755, 0B1-0B8, Bellingham, WA, USA http://dx.doi.org/DOI:10.1117/12.735803 (2007).
[20] G. Bonifazi, S. Serranti A. Bonoli, A. Dall'Ara, in C.A. Brebbia, M. Neophytou, E. Beriatos et al. (Eds.), Sustainable development and planning IV, Book Series: WIT Transactions on Ecology and the Environment, 120, 885-894. WIT Press, Southampton, Boston, ISBN: 978-1-84564-422-2.
[21] S. Serranti, A. Gargiulo, G. Bonifazi, Waste Manage. 31 (2011) 2217–2227.
[22] S. Serranti, A. Gargiulo, G. Bonifazi, Resour. Conserv. Recycl. 61 (2012) 52–58.
[23] A. Del Fiore, M. Reverberi, A. Ricelli, F. Pinzari, S. Serranti, A.A. Fabbri, G. Bonifazi, C. Fanelli, Int. J. Food Microbiol. 144 (2010) 64–71.
[24] M.A. Shahin, S.J. Symons, Comput. Electron. Agric. 75 (2011) 107–112.
[25] C.B. Singh, D.S. Jayas, J. Paliwal, N.D.G. White, Comput. Electron. Agric. 73 (2010) 118–125.
[26] C.M. McGoverin, M. Manley, J. Near Infrered Spectrosc. 20 (2012) 529–535.
[27] G. ElMasry, A. Iqbal, D.-W. Sun, P. Allen, P. Ward, J. Food Eng. 103 (2011) 333–344.
[28] M. Otto, Chemometrics, Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, New York, 1999.
[29] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[30] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627–1639.
[31] H. Martens, M. Høy, B.M. Wise, R. Bro, P.B. Brockhoff, J. Chemometr. 17 (3) (2003) 153–165.
[32] S. Wold, K. Esbensen, P. Geladi, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.
[33] M. Barker, W. Rayens, J. Chemometr. 17 (2003) 166–173.
[34] S. Wold, H. Martens, H. Wold, in: A. Ruhe, B. Kagstrom, (Eds.). Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad. Springer-Verlag, Heidelberg, Germany, 1983, pp. 286–293.
[35] R. Gosselin, D. Rodrigue, C. Duchesne, Chemom. Intell. Lab. Syst. 100 (2010) 12–21.
[36] S. Wold, E. Johansson, M. Cocchi, in: H. Kubinyi (Ed.), Drug Design: Theory, Methods and Applications, Escom Science Publishers, Leiden, The Netherlands, 1993, pp. 523–550.
[37] B. Moller, L. Munch, in: D.-W. Sun (Ed.), Infrared Spectroscopy for Food Quality Analysis and Control, Academic Press, USA, 2009, pp. 275–319.

Silvia Serranti conceived the work, defined the experimental set up, selected the procedures and the algorithms to apply, analyze the data and wrote the work. Eng. Daniela Cesare performed the analysis and contributes to evaluate the results. Dr. Federico Marini assisted in the data analysis step and contributed to the final writing of the paper. Prof. Giuseppe Bonifazi contributed to results evaluation and paper final writing.