Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors

Journal:	*Analytical and Bioanalytical Chemistry*
Manuscript ID	ABC-00307-2016.R1
Type of Paper:	Research Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Calvini, Rosalba; University of Modena and Reggio Emilia, Department of Life Sciences; University of Modena and Reggio Emilia, Interdepartmental Research Centre BIOGEST-SITEIA Foca, Giorgia; University of Modena and Reggio Emilia, Department of Life Sciences ; University of Modena and Reggio Emilia, Interdepartmental Research Centre BIOGEST-SITEIA Ulrici, Alessandro; University of Modena and Reggio Emilia, Department of Life Sciences; University of Modena and Reggio Emilia, Interdepartmental Research Centre BIOGEST-SITEIA
Keywords:	Hyperspectral imaging, data dimensionality reduction, fast exploration, multivariate classification, data fusion, green coffee

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA
Dipartimento di Scienze della Vita

Reggio Emilia, May 27, 2016

Dear Dr. Philippe Garrigues,

please find enclosed a copy of the revised version of manuscript ABC-00307-2016, that we would like to be considered for publication in the special issue of Analytical and Bioanalytical Chemistry "Chemical Sensing: from New Materials to in vivo Applications".

The manuscript has been deeply revised following the useful suggestions of the Referees, that we would like to thank very much for their precious comments.

We have addressed all the comments made by Referees A, B and C as indicated in the detailed list of the responses to the Referees questions that is enclosed. In the revised manuscript, all changes specific to Referee comments have been highlighted in red.

We are grateful to the Referees for their critical comments and useful suggestions that have helped us to improve our paper considerably.

The article is original, unpublished and not being considered for publication elsewhere.

Kind regards,

Alessandro Ulrici

Prof. Alessandro Ulrici
Dipartimento di Scienze della Vita
Università degli Studi di Modena e Reggio Emilia
Padiglione Besta - Via Amendola, 2 - 42122 Reggio Emilia - Italy
Phone +39 0522 522043
Fax +39 0522 522027
email: alessandro.ulrici@unimore.it

DATA DIMENSIONALITY REDUCTION AND DATA FUSION FOR FAST CHARACTERIZATION OF GREEN COFFEE SAMPLES USING HYPERSPECTRAL SENSORS

Rosalba Calvini, Giorgia Foca, Alessandro Ulrici*

Department of Life Sciences and Interdepartmental Research Centre BIOGEST-SITEIA, University of Modena and Reggio Emilia

* Corresponding author: Alessandro Ulrici, Department of Life Sciences, University of Modena and Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy. Phone: +39 0522 522043. Fax: +39 0522 522027. E-mail: alessandro.ulrici@unimore.it

ABSTRACT

Hyperspectral sensors represent a powerful tool for chemical mapping of solid-state samples, since they provide spectral information localized in the image domain in very short times and without the need of sample pretreatment. However, due to the large data size of each hyperspectral image, data dimensionality reduction (DR) is necessary in order to develop hyperspectral sensors for real-time monitoring of large sets of samples with different characteristics. In particular, in this work we focused on DR methods to convert the three-dimensional data array corresponding to each hyperspectral image into a one-dimensional signal (1D-DR), which retains spectral and/or spatial information. In this way, large datasets of hyperspectral images can be converted into matrices of signals, which in turn can be easily processed using suitable multivariate statistical methods. Obviously, different 1D-DR methods highlight different aspects of the hyperspectral image dataset. Therefore, in order to investigate their advantages and disadvantages, in this work we compared three different 1D-DR methods: average spectrum (AS), single space hyperspectrogram (SSH) and common space hyperspectrogram (CSH). In particular, we have considered 370 NIR-hyperspectral images of a set of green coffee samples, and the three 1D-DR methods were tested for their effectiveness in sensor fault detection, data structure exploration and sample classification according to coffee variety and to coffee processing method. Principal Component Analysis and Partial Least Squares-Discriminant Analysis were used to compare the three separate DR methods. Furthermore, low-level and mid-level data fusion was also employed to test the advantages of using AS, SSH and CSH altogether.

KEYWORDS

Hyperspectral imaging; data dimensionality reduction; fast exploration; multivariate classification; data fusion; green coffee.

1 INTRODUCTION

The use of hyperspectral imaging (HSI) is rapidly emerging in the field of analytical chemistry [1], since it allows to collect simultaneously and in short times spectral information localized in the spatial domain of a sample. Compared with conventional spectroscopic methods, HSI provides additional information by way of the acquisition of spatially resolved spectra, combining the advantages of imaging techniques with the attributes of spectroscopic measurements. This allows to obtain chemical maps of the analysed sample, i.e., false colour images where the different constituents on the sample surface are depicted with different colours, depending on their chemical composition. Therefore, the possibility to acquire chemical maps of the analysed samples in short times and without the need of sample pretreatment makes hyperspectral sensing a powerful tool for real-time monitoring of solid-state samples. However, for practical applications it is often necessary to account for the variability of a high number of representative samples, which implies the need to deal with large sets of hyperspectral images. Since each hyperspectral image consists in an extremely high number of spectra (generally tens of thousands), dealing with datasets containing hundreds of images requires efficient data handling strategies.

The data array corresponding to each hyperspectral image is often referred to as *hypercube*, given the three-dimensional nature of the hyperspectral data, with two spatial dimensions (y pixel rows and x pixel columns) and one spectral dimension (λ wavelengths) [2-5]. The hypercube can be therefore considered both as a series of spectrally resolved images, where each image is taken at a given wavelength, and as a series of spatially resolved spectra, where each one of the n ($= x \times y$) pixels of the image corresponds to a spectrum.

The major limitations to the development of robust and efficient predictive models based on hyperspectral sensors, that could be used for fast and continuous monitoring purposes in the industry, are due to the long times and to the high computational loads needed to efficiently extract the useful information from a high number of hyperspectral images altogether [6, 7]. In fact, when a model (e.g., a calibration or a classification model) must be developed considering simultaneously a lot of images, including all the single pixel spectra of all the images requires long computational times or could be even unfeasible. This drawback is particularly evident in the elaboration of datasets composed of tens up to thousands of hypercubes, and partly undermines the advantage of HSI to provide data extremely rich in information. Moreover, also for predictive purposes, when each image represents the properties of a specific sample (including both its average properties and its inner variability), it could be useful to obtain a direct estimate of the sample on the whole, before focusing on the single pixels.

To overcome this problem, the use of proper data dimensionality reduction (DR) methods is often mandatory. Among the various approaches used to perform DR, the simplest but still effective ones are those which consist in converting the three-dimensional hypercube data into a one-dimensional signal (1D-DR), which retains at least a significant part of the useful information pieces contained in each image. In this manner, the size of the dataset is drastically reduced; furthermore, the fast and simultaneous comparison at the image-level of a high number of hyperspectral images is possible. In fact, the whole dataset of hyperspectral images is converted into matrices of signals, which in turn can be easily processed by means of suitable multivariate statistical methods, like e.g., Principal Component Analysis (PCA) or Partial Least Squares – Discriminant Analysis (PLS-DA). In this manner, it is possible to visualize the structure of the whole dataset, highlighting clusters of similar samples, to develop classification rules and to detect outlier images, which could be then

further inspected in detail in order to reveal the reason for their anomalous behaviour, such as presence of defects within the analysed sample, contaminations or instrumental faults. A schematic overview of this approach is reported in Figure 1.

Depending on the specific 1D-DR method, different pieces of information of the hyperspectral image can be retained: the signal can contain only spectral information, only spatial information, or both. Frequently, reduction of hyperspectral images to one-dimensional signals is achieved by calculating the average spectrum from the whole image or from specific regions of interest (ROIs) selected within the image [8, 9]. In the case of homogeneous samples this strategy is effective, but it does not work well when the identification of spatially localized features within the image scene is needed, since average spectra (AS) do not account for spatial information.

Recently, following an approach that was previously developed by some of us for the analysis of datasets of RGB images [10-12], we proposed a 1D-DR method, named *hyperspectrogram* [6, 13], to condense both spatial and spectral information of hyperspectral images. Basically, hyperspectrograms are one-dimensional signals built by merging in sequence the frequency distribution curves of pixel-related features obtained from a PCA model calculated separately for each hyperspectral image. Therefore, spatial information is codified through the use of frequency distribution vectors of pixel-related features. Moreover, by adding at the end of the signal the PCA loading vectors, the hyperspectrogram codifies also the most relevant spectral features of the hypercube data. Since hyperspectrograms are generated from the outcome of a PCA model calculated on each separate hypercube, they reflect both spatial and spectral variability within each single hyperspectral image; for this reason, from here onwards they will be referred to as *Single Space Hyperspectrograms* (SSH).

Another approach based on the use of frequency distribution curves of pixel-related features for the discrimination between classes of objects (identified by the corresponding ROIs) within hyperspectral images has been recently proposed by Kucheryavsky [14]. In this work Kucheryavsky proposed to calculate a PCA model from a subset of images with representative objects from each considered class, and to use the frequency distribution curves of the score values of each principal component to build a feature vector for each object contained in the images. Following this idea, in the present paper we also introduce a novel version of hyperspectrograms, which in this case are based on a common PCA model, i.e., a model calculated on the whole dataset of hyperspectral images. For this reason, from here onwards these signals will be referred to as *Common Space Hyperspectrograms* (CSH). For each image, CSH are built by merging in sequence the frequency distribution curves of the corresponding pixel-related features. Obviously, in this case the loading vectors are not included into the signal, since they are the same for all the images. Therefore, each CSH codifies the spatial information of the corresponding hyperspectral image, while the spectral information is not considered explicitly.

Similarly to average image spectra, the main aim of SSH and CSH is therefore to provide the user with an additional tool for fast exploration of large datasets of hyperspectral images: a fast overview of the whole dataset can be made by applying multivariate explorative techniques, such as PCA, to matrices of AS, SSH or CSH, in which each signal represents a sort of fingerprint of the imaged sample. This inspection is performed at the image level (i.e., considering each image is a single object) and, in the case of hyperspectrograms, it allows considering aspects related to spatial variability of each image, such as local sample defects or instrumental

faults. The exploration of the whole dataset of hyperspectral images could allow for example to find groups of similar images, that can be subsequently inspected more in-depth at the pixel level (i.e., considering each pixel as a separate object) by means of more refined pixel-oriented techniques, such as Multivariate Image Analysis (MIA) [5, 15, 16] or Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [17]. Additionally, the matrices of AS, SSH and CSH can be also used for the development of multivariate classification or calibration models.

Therefore, it is important to emphasize that hyperspectrograms, and more in general 1D-DR methods, do not represent an alternative, but an additional way to analyse hyperspectral images, with respect to the "standard" pixel-based techniques, such as MIA. Indeed, MIA approach considers individual pixels as samples rather than the whole image, and thus this method does not allow to investigate the structure of the entire dataset of hyperspectral images. Conversely, hyperspectrograms allow considering each image as a sample, preserving at the same time spatial-related information. In this manner, a fast overview of the whole dataset can be easily achieved; this approach is particularly convenient when dealing with datasets composed of tens up to thousands of images, for example to evaluate the representativeness of a training set or differences between specific groups of images. As a matter of fact, in this situation performing MIA on each single image would imply long computational times and, furthermore, it would not allow to gain a global overview of the whole dataset.

In the present work, 1D-DR of hyperspectral images acquired in the near infrared (NIR) range (900-1700 nm) has been used for the classification of green coffee samples in *Arabica* (*Coffea arabica*) and *Robusta* (*Coffea canephora*) varieties, which are the two main species used for the preparation of commercial coffee beverages. Due to its better taste and aroma, *Arabica* coffee is of higher quality than *Robusta* coffee, but it is more difficult to grow, due to its lower resistance to plant diseases, and therefore it is more expensive [18]. For this reason, it is important to correctly discriminate between the two coffee varieties in order to prevent food frauds. Coffee varietal differentiation based on conventional NIR spectroscopy has been widely investigated in the last 20 years [19-22], proving that the NIR spectrum contains the information useful to discriminate between the two coffee varieties. However, conventional NIR spectroscopy does not allow to obtain chemical maps of the analysed samples, and therefore to detect spatially localized features such as impurities.

Moreover, the green coffee samples were also classified on the basis of the different processing methods that were used to separate the fruit from the seed. Some studies, in fact, highlighted that the kind of post-harvest treatment affects the contents of fructose and glucose in the coffee beans as well as the γ-aminobutyric acid content [23, 24].

All the hyperspectral images that were acquired on the green coffee samples were converted into the corresponding AS, SSH and CSH. The three matrices of signals were then analysed separately each other; firstly, the data structure was explored by means of PCA, and then classification rules were calculated using PLS-DA.

Finally, in order to evaluate the possibility to take advantage from considering the three kinds of signals altogether, data fusion strategies were used to combine the information brought by AS, SSH and CSH. Data fusion is generally applied in order to jointly analyse data collected from multiple sources and integrate the information conveyed from different analytical measurements [25, 26]. In our case, data fusion was used in order to take advantage of the complementary information carried by the different signals obtained from the original hyperspectral images. In particular, two different data fusion strategies were considered: low-level data fusion, that consisted in merging together the different signals, and mid-level data fusion, that consisted in combining

features extracted from the original signals. The classification results obtained from the three kinds of signals (AS, CSH and SSH) separately and from fused data were compared and discussed.

2 EXPERIMENTAL

2.1 *Samples*

A local roasting company provided aliquots of green coffee beans properly sampled from 31 batches, coming from 8 different geographical areas, and representative of the various types and amounts of green coffee used by the roasting company in the considered time span. Each aliquot consisted of about 500 g of beans that were sampled in order to be as representative as possible of the corresponding batch. The batches belonged to *Arabica* and *Robusta* varieties, and were subjected to different processing methods to remove the pulp of the coffee cherries from the seed. In particular, the *dry method* is used to produce *Natural* coffee, the *wet method* is used to produce *Washed* coffee, while an intermediate processing method, referred to as *polishing method*, is used to produce *Polished* coffee [27]. The aliquots were delivered to the laboratory in 4 different days covering a period of about 5 months (April 14th; June 6th; July 3rd; August 28th), and were stored at room temperature in sealed packages until the analysis.

Two replicate image-acquisition sessions were done. In the first session, immediately after delivery, three different samples of 70 g of randomly selected beans were taken from each package, and for each sample two repeated images were acquired considering different arrangements of the beans: the first image was acquired on piled beans, and the second one after shuffling and then levelling the beans in one layer. After image acquisition the samples were stored separately, and the original package containing the remainder green coffee beans of each batch was sealed again and stored at room temperature. The same procedure was repeated in the second image-acquisition session, few days after completion of the first session, repeating the same image acquisition procedure on three further 70 g green coffee samples for each batch. In both sessions, the batches were analysed in random order. Therefore, 12 hyperspectral images were collected on the whole for each batch, corresponding to 6 samples × 2 images (acquired on piled and levelled coffee beans, respectively), leading to a dataset composed of 372 hyperspectral images (31 batches × 12 images). During the following image processing phase 2 image files turned out to be corrupted, corresponding to one *Robusta Natural* sample and one *Robusta Washed* sample. Therefore the dataset was finally composed of 370 hyperspectral images. A detailed description of the considered batches is given in Table 1.

2.2 *Hyperspectral image acquisition*

The hyperspectral images were acquired using a desktop NIR Spectral Scanner (DV Optic) embedding a Specim Imspector N17E reflectance imaging spectrometer coupled to a Xenics Xeva-1.7-320 camera (320 × 256 pixels). The hyperspectral system is equipped with Specim Oles 31 f/2.0 optical lens and is characterized by a spectral distance of pixels equal to 3.26 nm, a maximum of spatial bending equal to 0.125 nm, straylight minor than 0.8% and a percentage of bad pixels equal to 0.37 % (automatically corrected through interpolation by the calibration pack furnished by Xeva). The system covers the 900-1700 nm spectral range with a spectral resolution of 5 nm.

Due to the low S/N ratio of the spectral region around 900 nm, only the 150 spectral channels between 955 nm and 1700 nm were considered in the present work. All the images were acquired using a black silicon carbide sandpaper sheet as background, which has a very low and constant reflectance spectrum [28]. In addition, each image was acquired in a manner to include in the scene a white ceramic tile 99 % reflectance standard reference and two ceramic tiles with average reflectance values equal to 89 % and 46 %, respectively. A picture and schematic representation of the acquisition system are reported in Figure S1, provided as Online Resource.

The raw data were converted into reflectance values by applying the instrument calibration procedure based on the reflectance standard reference and on the estimate of the dark current [29]. Furthermore, in order to reduce the variability among images over time, an additional internal calibration was performed [6, 30], based on the average reflectance values of the reflectance standard reference, of the two ceramic tiles and of the black silicon carbide sandpaper.

Finally, from each image the pixels related to the black sandpaper background were removed. In particular, a preliminary evaluation of a subset of images led to the identification of the most discriminant wavelength by maximizing the Fisher ratio between background spectra and sample spectra: this allowed to identify as background and remove all the pixels below the threshold value of 0.1 reflectance units measured at 1050 nm. The spectra retained after background elimination of one image of *Arabica* coffee and one image of *Robusta* coffee are reported in Figure S2, provided as Online Resource. The size of the final dataset of hyperspectral images was equal to 10.5 GB.

2.3 Data dimensionality reduction of hyperspectral images to one-dimensional signals

Three methods were used in order to reduce the data dimensionality of each hypercube to a one-dimensional signal (1D-DR), i.e.:

1) Average Spectra (AS), that codify only spectral information contained in each hyperspectral image;
2) Single Space Hyperspectrograms (SSH), that codify both spectral and spatial information contained in each hyperspectral image;
3) Common Space Hyperspectrograms (CSH), that codify only spatial information contained in each hyperspectral image.

Before the conversion into the three kinds of signals, the hyperspectral images were split into training set images, used for the creation of the matrices of training set signals, and into two separate subsets of test set images. The training set images correspond to the first two deliveries, while the two separate subsets of test set images correspond to day-3 and day-4 deliveries. This subdivision was done in order to mimic control procedures in the industrial plant, where models built on historical data are used to predict new incoming batches.

The plots of the three different kinds of signals obtained from the original images are reported within the schematic overview of Figure 1 (AS matrix, SSH matrix and CSH matrix), to allow a visual comparison of their aspect. In this figure, the signals of *Arabica* and *Robusta* classes are represented with different colours.

The detailed procedures used to obtain the three types of signals are described in the following sections.

2.3.1 *Conversion of images into Average Spectra*

After background elimination, from each image belonging to the training set images and to the two subsets of test set images the average spectrum was calculated by unfolding the hypercube to a matrix of pixel spectra with size {$n \times 150$}, where n is the number of image pixels retained after background removal, and 150 is the number of spectral channels. The average spectrum was then calculated as the arithmetic mean of the n values for each spectral channel. Therefore, the hyperspectral image dataset was converted into a matrix of 370 average spectra, each one containing 150 variables. The whole dataset of AS, stored in a file with size equal to 408 KB, consisted in a training set with size {203×150} and in two test sets, i.e., day-3 test set with size {83×150} and day-4 test set with size {84×150}. The average spectra belonging to the entire AS dataset are reported in Figure S3, provided as Online Resource.

2.3.2 *Conversion of images into Single Space Hyperspectrograms*

The procedure used to convert hyperspectral images into SSH is schematically depicted in Figure 2a. As mentioned above, the idea behind hyperspectrograms is to compress the potentially useful information contained in each hyperspectral image into a signal, by merging together quantities derived by a PCA model calculated on the considered image. A detailed description of the procedure used to calculate SSH is reported in the previously published articles [6, 13]; here below a summary of the procedure used in this work is reported.

For each image, the calculation of the corresponding SSH involved the following steps:

1. unfolding of the three-dimensional hypercube into a two-dimensional matrix containing as many rows as the pixels retained after background elimination and as many columns as the number of wavelengths;
2. calculation of a PCA model on mean centred spectra considering 3 PCs and storage of the corresponding score, Q-residuals, Hotelling T^2 and loading vectors. The number of PCs bringing useful information (3) and the most appropriate spectra preprocessing method (mean centering) were defined on the basis of a preliminary evaluation made by PCA on a restricted number of representative images;
3. calculation of frequency distribution curves for each score vector and for the Q residuals and T^2 vectors, considering a common scaling range for each vector equal to the global minimum and global maximum of the outputs of the PCA models obtained from the training set images and considering a number of bins equal to the number of spectral variables (150);
4. normalization of each frequency distribution curve by the number of pixels retained after background elimination;
5. creation of the SSH by joining in sequence the frequency distribution curves of the scores vectors, of the Q residual vector and of the T^2 vector, and finally adding the loading vectors; the length of the resulting signal is therefore equal to 1200, i.e.: (5 frequency distribution curves + 3 loading vectors) × 150 variables each.

The whole dataset of SSH, stored in a file with size equal to 2.30 MB, consisted in a training set with size {203×1200} and in two test sets, i.e., day-3 test set with size {83×1200} and day-4 test set with size {84×1200}.

2.3.3 *Conversion of images into Common Space Hyperspectrograms*

The procedure used to convert hyperspectral images into CSH is schematically depicted in Figure 2b. In this alternative approach, a global PCA model was computed using the images of the training set. The calculation of CSH involved the following steps:

1. unfolding of all the three-dimensional hypercubes into two-dimensional matrices containing as many rows as the pixels retained after background elimination and as many columns as the number of wavelengths;
2. (optional) row-preprocessing of the unfolded hypercubes using proper methods. In this case, SNV was chosen after a preliminary evaluation made by PCA on a limited number of sample images, and also coherently with the row-preprocessing used for the AS dataset (*see* section 2.4.1).
3. mean centering of the unfolded (and row-preprocessed) hypercubes according to the grand mean, i.e., to the mean calculated for each spectral channel on all the retained pixel spectra belonging to the training set images. The grand mean was calculated as the weighted arithmetic mean of the training set AS matrix, considering the number of pixels retained after background elimination as weights;
4. calculation of the kernel variance–covariance matrix of the whole dataset, Z [31]. The elements of Z are the covariances between the spectral channels, therefore the size of Z is equal to {150 × 150}; in order to make the calculation of Z computationally feasible, the variance–covariance matrices were calculated separately for each image, and then they were summed entrywise to obtain Z;
5. decomposition of Z by Singular Value Decomposition (SVD) to obtain the loading vectors of the common PC space (in this case, 5 PCs were considered, *see* below);
6. computation of the score, Q residuals and Hotelling T^2 vectors of each image by projecting the images onto the common PC space;
7. calculation of frequency distribution curves for each score vector and for the Q residuals and Hotelling T^2 vectors, considering a number of bins equal to 150;
8. normalization of each frequency distribution curve by the number of pixels retained after background elimination;
9. creation of the CSH by joining in sequence the frequency distribution curves of the score vectors, of the Q residual vector and of the Hotelling T^2 vector; the length of the resulting signal is therefore equal to 1050, i.e., 7 frequency distribution curves × 150 variables each.

As far as the images of the two test sets are concerned, after row-preprocessing the spectra as defined in step 2 and subtracting the grand mean calculated in step 3, the corresponding CSH were calculated following steps 6-9.

The whole dataset of CSH, stored in a file with size equal to 1.13 MB, consisted in a training set with size {203 × 1050} and in two test sets, i.e., day-3 test set with size {83 × 1050} and day-4 test set with size {84 × 1050}.

It has to be noticed that SSH and CSH were calculated considering a different number of PCs (3 and 5, respectively). Indeed, since CSH approach is based on a global PCA model, i.e., on a model including all the spectra of all the images of the training set, it accounts for more sources of variance than SSH, which conversely considers only the pixels variability within each single image. In particular, for CSH the first PCs usually

account for between-images variability and, therefore, it is generally convenient to retain a higher number of PCs respect to SSH. However, including in both approaches also the frequency distribution curve of the Q residuals ensures that all the information which is potentially useful for a given problem is somehow considered.

2.3.4 Computation time

For an exhaustive comparison of the three 1D-DR methods, it is worth mentioning the computation time required for the conversion of images into signals. In fact, even if the calculation of the hyperspectrograms goes through several steps, it has to be underlined that all these steps are completely automated and much faster than performing the analysis of single images or of groups of them at the pixel level using MIA.

Running the algorithms in Matlab ver. 7.12, and using a personal computer with Microsoft Windows 7–64 bit OS, equipped with an Intel Core i7-2600 CPU @ 3.40 GHz processor and 4.00 GB RAM, the average computation time of each AS was equal to 1.0 s, the corresponding time required to calculate a SSH resulted equal to 1.2 s and the time required to calculate a CSH was equal to 2.3 s. Therefore, the calculation of SSH is nearly as fast as the calculation of AS, while the average computation time of CSH is about twice. However, the computation time equal to 2.3 s for CSH is referred to the training set signals; in this case, the procedure involves the calculation of a common PCA model for all the training set images. Once the common set of loading vectors is obtained, the computation time of the CSH for a test image is the same as for an AS.

As a matter of fact, since our dataset is composed of 370 hyperspectral images, a direct comparison with "standard" multivariate image analysis methods working at the pixel level (i.e., considering all pixel spectra of each hyperspectral image as separate objects) would be unfeasible. For example, considering the same 203 training set images, each one containing on average 25000 pixels after background elimination, working at the pixel level would require to calculate a PLS-DA classification model on a dataset with more than 5 million (= 203 × 25000) of objects and 150 variables (spectral channels). It would be impossible, or at least very difficult, to manage directly such a dataset size by commonly used hardware and software tools.

2.4 Data analysis on signals

2.4.1 Data exploration by PCA

The exploration of the structure of the three datasets composed by the AS, SSH and CSH signals was done by PCA. Various preprocessing methods were tested on AS dataset to remove uninformative variation and to enhance the clustering of the objects belonging to the different classes, including mean centering, first and second order derivatives and SNV. SNV followed by mean centering resulted to be the most effective preprocessing for the AS dataset. Conversely, since CSH and SSH are signals obtained by merging in sequence frequency distribution curves of quantities derived from PCA models, the use of row preprocessing methods is not adequate and, for this reason, both the SSH and CSH datasets were only mean centered.

It must be underlined that SNV was also applied as row preprocessing method of the spectra before the calculation of the global PCA model for the construction of CSH (as described Section 2.3.3). In fact, the use of SNV resulted useful for the analysis of AS dataset and for the calculation of CSH, since both of these approaches imply the direct comparison between spectra of different hypercubes (average spectra for AS, all the spectra

altogether for CSH). Conversely, to create the SSH, the spectra were preprocessed using only mean centering, since in this case the spectra of each image were considered separately from those of the other images.

2.4.2 *Classification by PLS-DA*

Partial Least Squares – Discriminant Analysis (PLS-DA) was used as classification method, in order to discriminate the green coffee samples based on their variety, i.e., *Arabica* vs. *Robusta*, and on the different processing methods used to remove the pulp from the seed, i.e. considering the *Natural*, *Washed*, *Polished* classes. PLS-DA classification rules were calculated for each one of the three datasets, i.e., AS, SSH and CSH, using the signal preprocessing methods defined in Section 2.4.1. The model dimensionality was chosen by minimizing the value of the Root Mean Square Error in Cross-Validation (RMSECV). In particular, a contiguous block cross-validation (CV) scheme with 17 deletion groups was considered, where each deletion group contained all the signals derived from the same batch.

Classification results were expressed in terms of True Positives (TP) rate, i.e., the percentage of objects of each class that were correctly assigned to the corresponding category, and of True Negatives (TN) rate, i.e., the percentage of objects belonging to other classes that were correctly rejected from the considered class. Furthermore, in order to gain an overall estimate of the performances of the different PLS-DA models, the classification results were also expressed in terms of Non-Error Rate (NER), i.e., the overall percentage of correctly assigned objects [32]. In particular, for each PLS-DA model, the TP, TN and NER values were calculated in calibration and in cross-validation for the training set, and in prediction for the day-3 test set (day-4 test set was not considered for the reasons specified below in Section 3.1). In order to evaluate the variables that mainly contributed to the classification rules, the Variable Importance in Projection (VIP) score plots of the PLS-DA models have been examined [33]. VIP scores are defined in a way that only those variables whose VIP score values are greater than 1 furnish a substantial contribution to the model.

2.4.3 *Data fusion*

Since AS, SSH and CSH bring different pieces of spectral-related (AS and SSH) and of spatial-related information (SSH and CSH), data fusion was also used in order to evaluate the possible advantages deriving from the synergistic use of the three kinds of signal in the calculation of PLS-DA models. In particular, two different data fusion strategies were considered, i.e., Low-Level (Low-L) and Mid-Level (Mid-L) data fusion [34].

Low-level data fusion consisted in merging in sequence the AS, SSH and CSH blocks. In this case, each block was first preprocessed separately as described in Section 2.4.1 and then it was block scaled, i.e., it was normalized to unit standard deviation. In this manner, while preserving the relative weights of the variables within each block, the three blocks contribute with equal weight to the calculation of the PLS-DA models. The resulting Low-L dataset, with size {370 × 2400} (i.e., 150 variables for AS + 1200 variables for SSH + 1050 variables for CSH), was subdivided in the same manner as for the single datasets, i.e., in a training set with size {203 × 2400} and day-3 test set with size {83 × 2400}, while day-4 test set was not considered for the reasons specified below in Section 3.1. Then, PLS-DA was applied to the Low-L fused data for the classification between *Arabica* and *Robusta* varieties, and for the classification between *Polished*, *Natural* and *Washed* categories.

Mid-level data fusion is based on combining features that were previously extracted from each block separately and then concatenating them into a single array, which can be subsequently used for the development of multivariate classification or regression models. These features can be either original variables identified by variable selection algorithms or latent variables, like e.g., PCA, PLS or PLS-DA scores [34-37]. In our case, Mid-L data fusion was performed by merging together the scores that were obtained from the PLS-DA models calculated on each single dataset of signals, as described in the previous section. Therefore, two different datasets were obtained from Mid-L data fusion: one for the classification between *Arabica* and *Robusta* varieties, and one for the classification between *Polished*, *Natural* and *Washed* categories. Also in this case, the Mid-L datasets were split following the same sample subdivision considered for the previous datasets. Then, PLS-DA was applied to Mid-L fused data; in this case, two preprocessing methods were tested, i.e., block-scaling and autoscaling.

All the functions used for the calculation and exploration of the hyperspectrograms were written in Matlab language (ver. 7.12, The Mathworks Inc., USA) and have been implemented into a Graphical User Interface (GUI) which is freely available upon request to the corresponding author. The PLS-DA classification rules were calculated using the PLS_Toolbox (ver. 7.8.2, Eigenvector Research Inc., USA).

3 RESULTS AND DISCUSSION

3.1 Explorative analysis by PCA

The PCA model calculated on AS training set data was found to have an optimal dimensionality equal to 3 PCs, accounting for about 98 % of total data variance. The score plot of the first two PCs is reported in Figure 3a, where the samples corresponding to *Arabica* and *Robusta* varieties are partially separate each other. The sub-clusters that are visible within each class correspond to different coffee batches. The corresponding loading vectors are reported in Figure 3b, which shows that the most relevant spectral variables involved in the separation between *Arabica* and *Robusta* varieties are related to the C-H second overtone and combination bands (at about 1100 nm, 1200 nm and 1350 nm) and to the O-H first overtone (1450 nm) [38, 39]. The samples resulted also partially clustered based on the *Natural*, *Washed* and *Polished* categories; in particular, the samples belonging to the *Robusta-Natural* batches were clustered together, like those of the *Robusta-Polished* batches (Figure 3a and Figure S4, provided as Online Resource).

For the SSH training set, a 4 PCs model was obtained, explaining about 94 % of the total data variance. *Arabica* and *Robusta* varieties were mainly distinguishable along PC4, as reported in the PC1 vs. PC4 scores plot of Figure 3c. The PC 4 loading vector reported in Figure 3d shows that, in this case, the relevant variables involved in the separation are related to the portions of SSH signals corresponding to the frequency distribution curves of scores (1-450) and Q residuals (451-600), and to the PC3 loading vectors (1050-1200). A further inspection revealed that SSH were able to extract features related to the different arrangement of the beans within the image scene; in fact, the images were clustered along PC2, reflecting the different arrangements of the coffee beans (levelled in one layer or piled) during image acquisition (Figure S5, provided as Online Resource). Furthermore, a partial separation of the *Polished* samples was observed along PC1 (Figure S6.a, provided as

Online Resource), while PC3 did not show remarkable trends, except for some variations within the images taken from one batch (cyan diamonds in Figure S6.b, provided as Online Resource).

Also for the CSH training set a 4 PCs model was obtained, explaining about 88 % of the total data variance. In Figure 3e, that reports the corresponding PC1 vs. PC4 scores plot, a quite clear distinction between *Arabica* and *Robusta* varieties is visible. The PC1 and PC4 loading vectors are reported in Figure 3f, which shows that the relevant variables in the separation of *Arabica* and *Robusta* samples are related to the portions of CSH signals corresponding to the frequency distribution curves of the scores (1-750), in particular of PC4 (601-750), and of the Hotelling T^2 vector (901-1050). For this dataset, PC2 and PC3 show the difference of three batches of *Washed* and *Natural Arabica* samples, delivered for analysis both on day-1 and day-2, with respect to all the other samples; also a partial separation of the cluster of *Polished* samples can be appreciated (Figure S7, provided as Online Resource).

For all the three types of signals the samples of day-3 and day-4 deliveries were then projected on the corresponding PCA models. In Figure 4, the reduced Q residuals (Q_r) and the reduced Hotelling T^2 values (T^2_r) are reported for AS (Figure 4a), SSH (Figure 4b) and CSH (Figure 4c) datasets. The values of Q_r and of T^2_r are calculated as the Q residuals and Hotelling T^2 values normalized by the corresponding 95 % confidence limit, in a way to allow direct comparison between the values calculated for different PCA models. Moreover, in order to better highlight the samples lying outside the confidence limit, the values of Q_r and of T^2_r have been reported on a logarithmic scale.

While all the objects of day-3 delivery are similar to the training set objects for all the datasets, many of the objects of day-4 delivery are outliers. In particular, the deviations are more marked for the PCA model calculated on the SSH dataset, where all the objects of day-4 delivery have Q_r values higher than the 95 % confidence limit. Based on these results, in order to reveal the cause of this anomalous behaviour, the images of day-4 delivery were then analysed more in depth at the pixel level by MIA, i.e., by calculating a PCA model on each single hyperspectral image. The score images obtained by MIA revealed the presence of a sort of striping. As an example, Figure 5a shows the PC3 score image of one sample of day-4 delivery, together with the NIR reflectance spectra corresponding to 4 specified points of the image (Fig. 5b). It can be noticed that, regardless of the offset, the reflectance spectra corresponding to the two points on the left in the score image (points 1 and 2) have a similar shape, which is different from the shape of the spectra corresponding to the points placed on the contiguous stripe of the score image (points 3 and 4).

The striping effect found on the images of day-4 delivery was unexpected, since it was not possible to notice this problem during image acquisition by simply visualizing the average intensity images. Based on the results reported in Figure 4 and Figure 5, a technical inspection of the instrument was performed, revealing that this problem was due to a fault of the detector cooling system. This finding indicates that AS gave models less sensitive to instrumental variability with respect to hyperspectrograms. On the other hand, instrument-related problems in the imaging system (such as variations with time or temperature) should be clearly identified. Since hyperspectrograms are more influenced by these sources of variability in the images, they may also provide an useful tool for monitoring the stability of the imaging system. In particular, among the three 1D-DR methods, SSH were the most sensitive to this instrumental fault, since they essentially reflect the spectral variability within each single image. For the same reasons, SSH could help to better highlight other within-image sources of variation, like the presence of local defects due for example to the presence of imperfect beans or impurities.

Since the images of day-4 delivery were found to be defective, they were not used in the following classification step.

3.2 *Classification results*

Table 2 reports the performances obtained for the discrimination between *Arabica* and *Robusta* using both the separate approaches and the fused data; the corresponding confusion matrices are reported in Table S1, provided as Online Resource. Considering the prediction of day-3 test set, common space hyperspectrograms gave the best results: the NER value reached 100 % with a model dimensionality lower than that of the other datasets. The better performance of CSH with respect to AS and SSH can be reasonably ascribed to the fact that each single common space hyperspectrogram is built in a way to consider both the within-image and the between-images spectral variability. In fact, while the frequency distribution curves composing the CSH account explicitly for the distribution of the pixel-spectra within each single image, the direction of the principal components along which the frequency distributions are calculated depends on the variability of all the spectra of all the images, i.e., it is strongly influenced by the between-images variability.

Focusing on the classification results of fused data, the performance in prediction of day-3 images always improved with respect to those of AS and SSH datasets, while it was equal or slightly worse with respect to CSH dataset alone. Among the different strategies adopted for data fusion, both block-scaled Low-L and block-scaled Mid-L datasets gave high classification results, with NER values for the prediction of day-3 test set equal to 99,8 % and 100 %, respectively.

Figure 6.a reports the Variable Importance in Projection (VIP) scores of the block-scaled Mid-L model, which shows that all the three blocks (AS, SSH and CSH) contributed to the discrimination between *Arabica* and *Robusta* coffee varieties. In particular, the first latent variable (LV1) scores of each block gave a significant contribution to the classification, and the highest value was obtained for LV1 scores from CSH. These results further confirm that CSH bring more useful information than AS and SSH for the discrimination between *Arabica* and *Robusta*. Conversely, SSH (both alone and in combination with AS and CSH) led to the worst classification performances. This is likely due to the fact that SSH are more suitable for the characterization of heterogeneous samples based on their inner variability, while both AS and CSH allow considering average properties of the images, which is the key aspect in the present task. Compared to AS, CSH retain also additional information related to spatial (within-image) variability, which contributed to their better classification performances.

Table 3 reports the results for the classification of *Natural*, *Washed* and *Polished* green coffee samples, obtained using both the separate approaches and the fused data; the corresponding confusion matrices are reported in Table S2, provided as Online Resource. Generally, the classification of coffee samples in the three categories corresponding to the different technological treatments used to remove the pulp from the seed resulted less trivial than discriminating the coffee varieties.

Considering the individual sets of signals, the *Polished* samples were classified satisfactorily using AS, which led to a TP rate equal to 91,7 % and a TN rate equal to 98,6 % for the prediction of day-3 test set samples. This is not surprising, since the PCA models have formerly evidenced a more clear clustering of this class of samples. On the other hand, for the day-3 samples belonging to the *Natural* and *Washed* classes, SSH and CSH

led to more satisfactory results with respect to AS. A comparison of the overall results (NER values) showed that CSH was the data dimensionality reduction method leading to the best results, which is reasonably due to the same reasons already discussed for the discrimination between *Arabica* and *Robusta* varieties.

After data fusion, the classification performances generally improved. The overall results highlight that also in this case the block-scaled Mid-L dataset gave the best results, leading to the highest NER value for the prediction of day-3 samples (90.4 %). In Figure 6.b the VIP scores of this PLS-DA model are reported. For the *Polished* class, VIP scores higher than 1 were obtained for the LV1 scores of each block; for the *Natural* class the main contribution was from LV2 scores of AS, from LV2 scores of SSH and mainly from LV2 and LV3 scores of CSH; for the *Washed* class the more significant variables were LV1 scores from AS and LV1, LV2 and LV3 scores from CSH. Therefore, in general the major contribution to the discrimination between the three classes was from CSH, further confirming the effectiveness of this approach.

Conversely to what observed for the discrimination between *Arabica* and *Robusta* varieties, mid-level data fusion helped to improve the classification results when dealing with a more complex classification issue such as the discrimination of the coffee samples into *Polished*, *Natural* and *Washed* categories. The better classification performances obtained with mid-level data fusion can be explained considering that, instead of the original AS/SSH/CSH variables, this approach makes use of the latent variables obtained by the PLS-DA models calculated on the single datasets, where the most part of information not pertinent to the discrimination issue is removed.

4 CONCLUSIONS

The present work described the application of data dimensionality reduction methods which can be used to automatically analyse large sets of samples with different characteristics by means of hyperspectral sensors. Hyperspectral images acquired on green coffee samples were converted into three different kinds of one-dimensional signals, namely Average Spectra (AS), Single Space Hyperspectrograms (SSH) and Common Space Hyperspectrograms (CSH). AS, SSH and CSH represent different data dimensionality reduction strategies, and highlight different aspects of the hyperspectral images. AS allow considering average properties of the images but, in this manner, the information related to spatial variability is lost. Therefore, this kind of approach is suitable for the analysis of homogeneous samples, but it might not give satisfactory results for the identification of spatially localized features, like, e.g., local sample defects. On the other hand, SSH are very effective in retaining spatial-related information of the images, but they are not the most suitable method when dealing with homogeneous samples or when considering properties that are common to the entire set of images. Finally, in the present study CSH approach is introduced, which can be seen as a sort of compromise between AS and SSH, since they take into account spatial related features, like SSH, and they also allow considering the average properties of each image, like AS. Therefore, the choice of the most proper data dimensionality reduction method strictly depends on the specific problem at hand.

These data dimensionality reduction methods were then used for fast exploration of the hyperspectral image dataset and for classification of the green coffee samples based on coffee variety and on raw beans processing method, with the aim of comparing their performances.

The conversion of images into signals allowed to drastically reduce the data size and to gain a fast overview of the structure of the whole dataset, which in turn permitted to detect the presence of defective images, due to an instrumental fault. Interestingly, the results of the explorative analysis performed with PCA showed that the hyperspectrogram approach (and in particular SSH) resulted to be more sensitive to the presence of time- and/or temperature-related variations of the images, which implies that they may be used as a tool for monitoring the stability of the imaging system.

The development of classification rules based on the three kinds of signals led generally to satisfactory results; in particular, CSH led to the best results for both classification tasks (i.e., coffee variety and raw beans processing method). Low-level and mid-level data fusion techniques were also employed, in order to evaluate the possible advantages deriving from the use of the information brought by all the three kinds of signal in the calculation of the classification rules. The best performances were always obtained using mid-level data fusion, which generally allowed to improve the classification results with respect to those obtained considering AS, CSH and SSH separately. In particular, the combined use of different 1D-DR approaches through data fusion resulted to be helpful to provide a significant improvement in the classification results when facing complex classification issues, such as the discrimination of the samples according to the raw beans processing method. Conversely, the discrimination of the samples based on coffee variety represented a simpler classification problem and, in this case, the hyperspectrogram approach was sufficient to obtain satisfactory results.

5 ACKNOWLEDGEMENTS

Dr. Luigi Bellucci of Caffè Molinari S.p.A roasting company is acknowledged for providing the samples and for technical support.

REFERENCES

1. Dale LM, Thewis A, Boudry C, Rotar I, Dardenne P, Baeten V, Pierna JAF. Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review. Appl Spectrosc Rev. 2013; 48(2):142–159.
2. Amigo JM. Practical issues of hyperspectral imaging analysis of solid dosage forms. Anal Bioanal Chem. 2010; 398:93–109.
3. Burger J, Gowen AA. Data handling in hyperspectral image analysis. Chemom Intell Lab Syst. 2011; 108:13–22.
4. Wu D, Sun D-W. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review – Part I: Fundamentals. Innov Food Sci Emerg. 2013; 19:1–14.
5. Amigo JM, Babamoradi H, Elcoroaristizabal S. Hyperspectral image analysis. A tutorial. Anal Chim Acta. 2015; 896:34–51.
6. Ferrari C, Foca G, Calvini R, Ulrici A. Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples. Chemom Intell Lab Syst. 2015; 146:108–119.
7. Gowen AA, Marini F, Esquerre C, O'Donnell CP, Downey G, Burger J. Time series hyperspectral chemical imaging data: challenges, solutions and applications. Analy Chim Acta. 2011; 705:272–282.
8. Gowen AA, O'Donnell CP, Taghizadeh M, Cullen PJ, Frias JM, Downey G. Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (Agaricus bisporus). J Chemom. 2008; 22(3-4):259–267.
9. Lindström SW, Nilsson D, Nordin A, Nordwaeger M, Olofsson I, Pommer L, Geladi P. Quality assurance of torrefied biomass using RGB, visual and near infrared (hyper) spectral image data. J Near Infrared Spec. 2014; 22(2):129–139.
10. Antonelli A, Cocchi M, Fava P, Foca G, Franchini GC, Manzini D, Ulrici A. Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. Anal Chim Acta. 2004; 515:3–13.
11. Foca G, Masino F, Antonelli A, Ulrici A. Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques. Anal Chim Acta. 2011; 706:238–24.
12. Ulrici A, Foca G, Ielo MC, Volpelli LA, Lo Fiego DP. Automated identification and visualization of food defects using RGB imaging: Application to the detection of red skin defect of raw hams. Innov Food Sci Emerg. 2012; 16:417–426.
13. Ferrari C, Foca G, Ulrici A. Handling large datasets of hyperspectral images: reducing data size without loss of useful information. Anal Chim Acta. 2013; 802:29–39.
14. Kucheryavskiy S. A new approach for discrimination of objects on hyperspectral images. Chemom Intell Lab Syst. 2013; 120:126–135.
15. Grahn H, Geladi P. Techniques and Applications of Hyperspectral Image Analysis. Chichester: John Wiley & Sons Ltd; 2007.
16. Prats-Montalbán JM, De Juan A, Ferrer A. Multivariate image analysis: A review with applications. Chemom Intell Lab Syst. 2011; 107(1):1–23.
17. Piqueras S, Burger J, Tauler R, De Juan A. Relevant aspects of quantification and sample heterogeneity in hyperspectral image resolution. Chemom Intell Lab Syst. 2012; 117:169–182.
18. Martín M, Pablos F, González A. Discrimination between arabica and robusta green coffee varieties according to their chemical composition. Talanta. 1998; 46:1259–1264.
19. Downey G, Briandet R, Wilson RH, Kemsley EK. Near- and Mid-Infrared spectroscopies in food authentication: coffee varietal identification. J Agric Food Chem. 1997; 45(11):4357–4361.
20. Esteban-Díez I, González-Sáiz JM, Pizarro C. An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS. Anal Chim Acta. 2004; 514(1):57–67.
21. Esteban-Díez I, González-Sáiz JM, Sáenz-González C, Pizarro C. Coffee varietal differentiation based on near infrared spectroscopy. Talanta. 2007; 71(1):221–229.
22. Buratti S, Sinelli N, Bertone E, Venturello A, Casiraghi E, Geobaldo F. Discrimination between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy, electronic nose and electronic tongue analysis. J Sci Food Agric. 2015; 95(11):2192–2200.
23. Knopp S, Bytof G, Selmar D. Influence of processing on the content of sugars in green Arabica coffee beans. Eur Food Res Technol. 2006; 223(2):195–201.
24. Bytof G, Knopp SE, Schieberle P. Influence of processing on the generation of γ-aminobutyric acid in green coffee beans. Eur Food Res Technol. 2005; 220(3-4):245–250.

25. Acar E, Rasmussen MA, Savorani F, Næs T, Bro R. Understanding data fusion within the framework of coupled matrix and tensor factorizations. Chemom Intell Lab Syst. 2013; 129:53–63.
26. Khaleghi B, Khamis A, Karray FO, Razavi SN. Multisensor data fusion: A review of the state-of-the-art. Inform Fusion. 2013; 14(1):28–44.
27. Bee S, Brando CHJ, Brumen G, Carvalhes N, Kolling-Speer I, Speer K, Suggi Liverani F, Teixeira AA, Teixeira R, Thomaziello RA, Viani R, Vitzthum OG. The Raw bean. In: Illy A, Viani R, editors. Espresso Coffee, 2nd Edition. Amsterdam: Elsevier Academic Press; 2005. pp. 91–96.
28. Burger J, Geladi P. Hyperspectral NIR image regression part II: Dataset preprocessing diagnostics. J Chemom. 2006; 20:106–119.
29. Burger J, Geladi P. Hyperspectral NIR image regression part I: Calibration and correction. J Chemom. 2005; 19:355–363.
30. Ulrici A, Serranti S, Ferrari C, Cesare D, Foca G, Bonifazi G. Efficient chemometric strategies for PET–PLA discrimination in recycling plants using hyperspectral imaging. Chemom Intell Lab Syst. 2013; 122:31–39.
31. Geladi P, Grahn H. Multivariate Image Analysis. Chichester: John Wiley & Sons Ltd; 1996.
32. Ballabio D., Todeschini R. Multivariate classification for qualitative analysis. Infrared Spectroscopy for Food Quality Analysis and Control, 2009, 83-104.
33. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst. 2005; 78:103-112.
34. Borràs E, Ferré J, Boqué R, Mestres M, Aceña L, Busto O. Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta, 2015, 891:1-14
35. Haware RV, Wright PR, Morris KR, Hamad ML. Data fusion of Fourier transform infrared spectra and powder X-ray diffraction patterns for pharmaceutical mixtures. J Pharm Biomed Anal. 2011; 56:944–949.
36. Bagnasco L, Cosulich ME, Speranza G, Medini L, Oliveri P, Lanteri S. Application of a voltammetric electronic tongue and near infrared spectroscopy for a rapid umami taste assessment. Food Chem. 2014; 157:421–428.
37. Biancolillo A, Bucci R, Magrì AL, Magrì AD, Marini F. Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. Anal Chim Acta. 2014; 820:23–31.
38. Shenk JS, Workman JJ, Westerhaus MO. Application of NIR Spectroscopy in Agricultural Products. In: Burns DA, Ciurczak EW, editors. Handbook of Near-Infrared Analysis, 3rd Edition. NewYork: Marcel Dekker; 2008. pp. 347-386.
39. Westad F, Schmidt A, Kermit M. Incorporating chemical band-assignment in near infrared spectroscopy regression models. J. Near Infrared Spectrosc. 2008; 16 (3):265-273.

FIGURE CAPTIONS

Figure 1. Schematic representation of the analysis of large datasets of hyperspectral images by means of 1D-DR methods.

Figure 2. Schematic description of the procedures used to convert images into single space hyperspectrograms (a) and into common space hyperspectrograms (b).

Figure 3. PC1 vs. PC2 scores plot of AS dataset (a) together with the corresponding loading vectors (b); PC1 vs. PC4 scores plot of SSH dataset (c) together with the corresponding loading vectors (d); PC1 vs. PC4 scores plot of CSH dataset (e) together with the corresponding loading vectors (f). *Arabica* coffee samples are represented in green colour and *Robusta* coffee samples are represented in red colour; *Polished* samples are represented with plus sign, *Natural* samples are represented with circles and *Washed* samples are represented with stars.

Figure 4. Q residuals and Hotelling T^2 plots of the PCA models calculated on AS (a), SSH (b) and CSH (c) datasets. Training set objects are represented with blue circles, test set objects of day-3 delivery with pink squares, and test set objects of day-4 delivery with cyan triangles. The dashed lines correspond to the 95 % confidence limit.

Figure 5. PC3 score image of a coffee sample acquired on day-4 presenting the striping effect (a) and reflectance spectra of different points of the same sample (b). The spectra are colored and numbered as the points to which they correspond in the score image.

Figure 6. VIP scores of PLS-DA models calculated with block-scaled Mid-L datasets for *Arabica/Robusta* classification (a) and *Polished/Natural/Washed* classification (b).

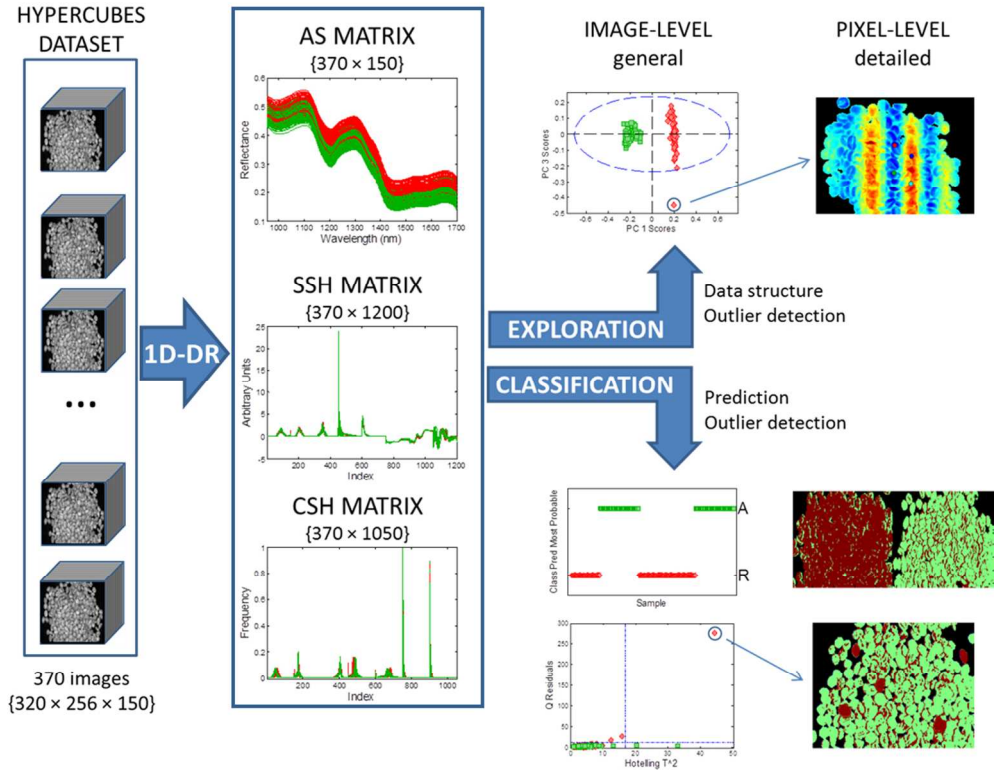

Figure 1. Schematic representation of the analysis of large datasets of hyperspectral images by means of 1D-DR methods.
302x234mm (96 x 96 DPI)

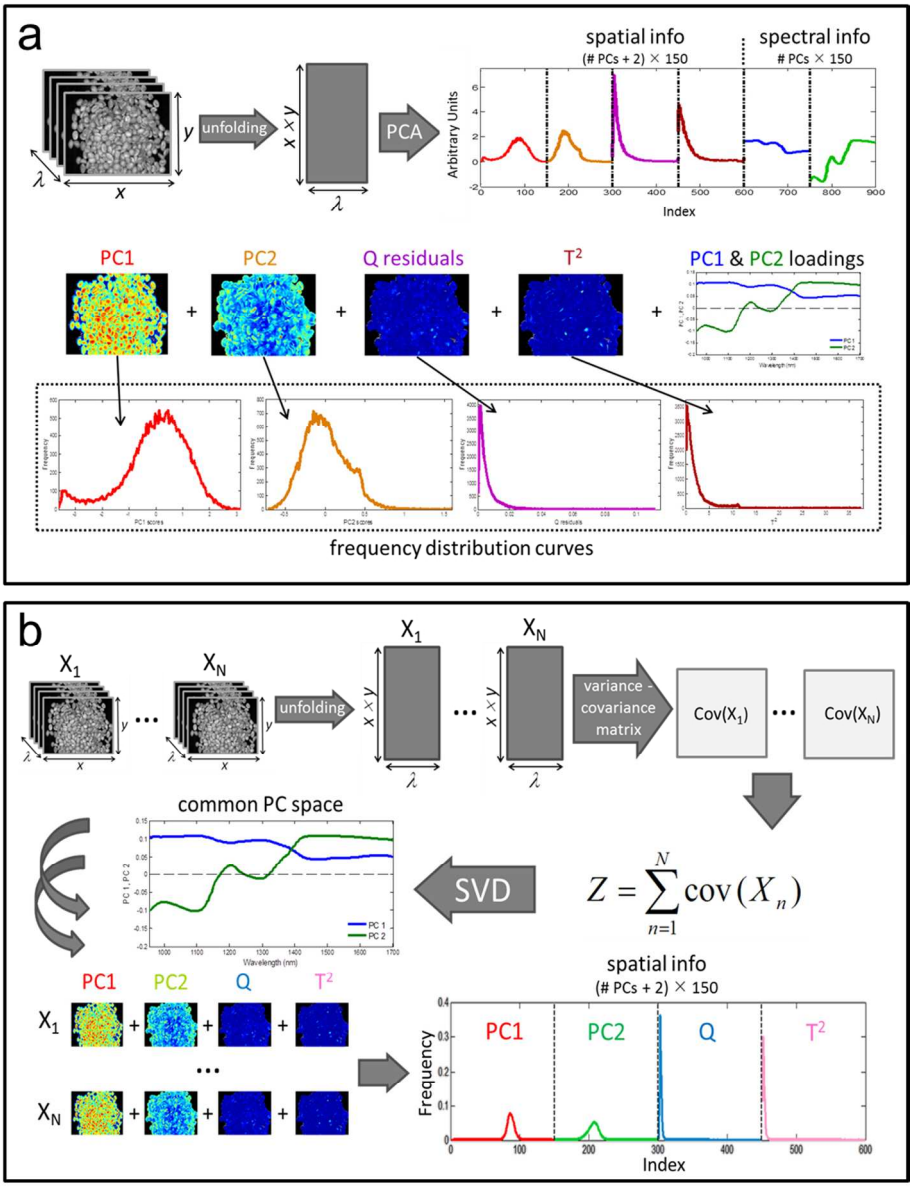

Figure 2. Schematic description of the procedures used to convert images into single space hyperspectrograms (a) and into common space hyperspectrograms (b).
299x389mm (96 x 96 DPI)

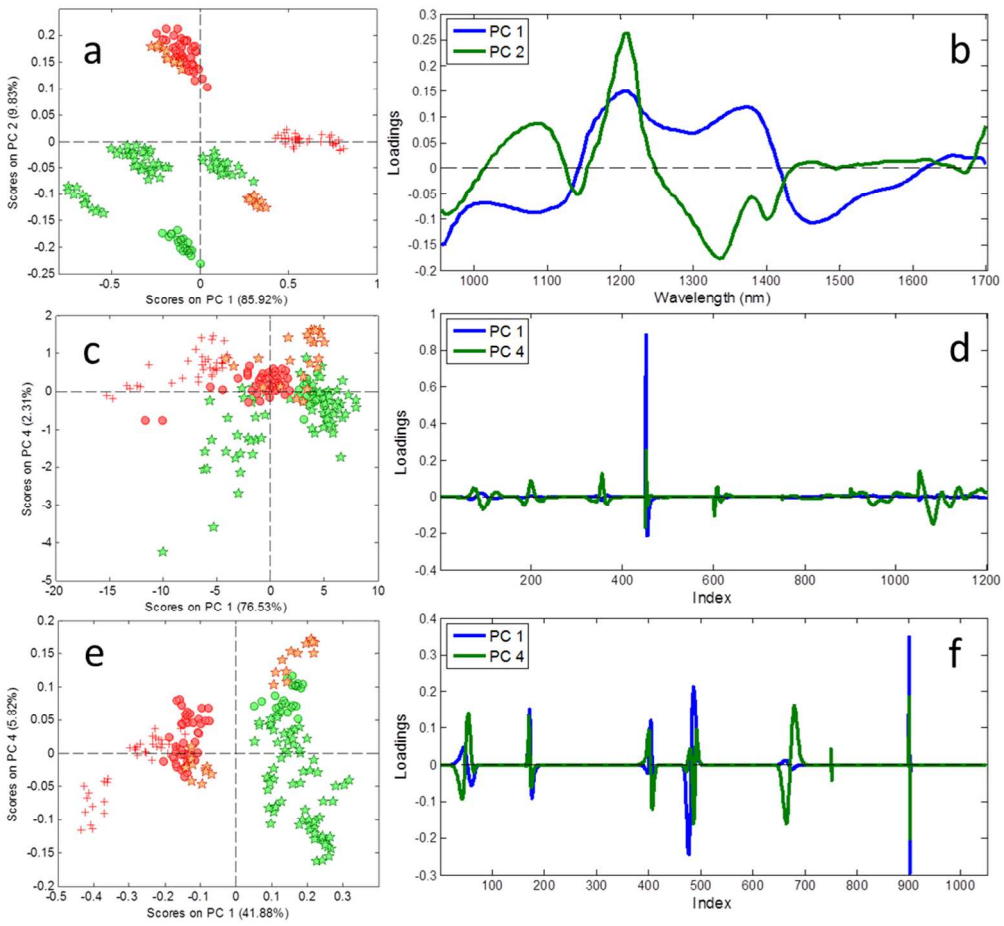

Figure 3. PC1 vs. PC2 scores plot of AS dataset (a) together with the corresponding loading vectors (b); PC1 vs. PC4 scores plot of SSH dataset (c) together with the corresponding loading vectors (d); PC1 vs. PC4 scores plot of CSH dataset (e) together with the corresponding loading vectors (f). Arabica coffee samples are represented in green colour and Robusta coffee samples are represented in red colour; Polished samples are represented with plus sign, Natural samples are represented with circles and Washed samples are represented with stars.
302x284mm (96 x 96 DPI)

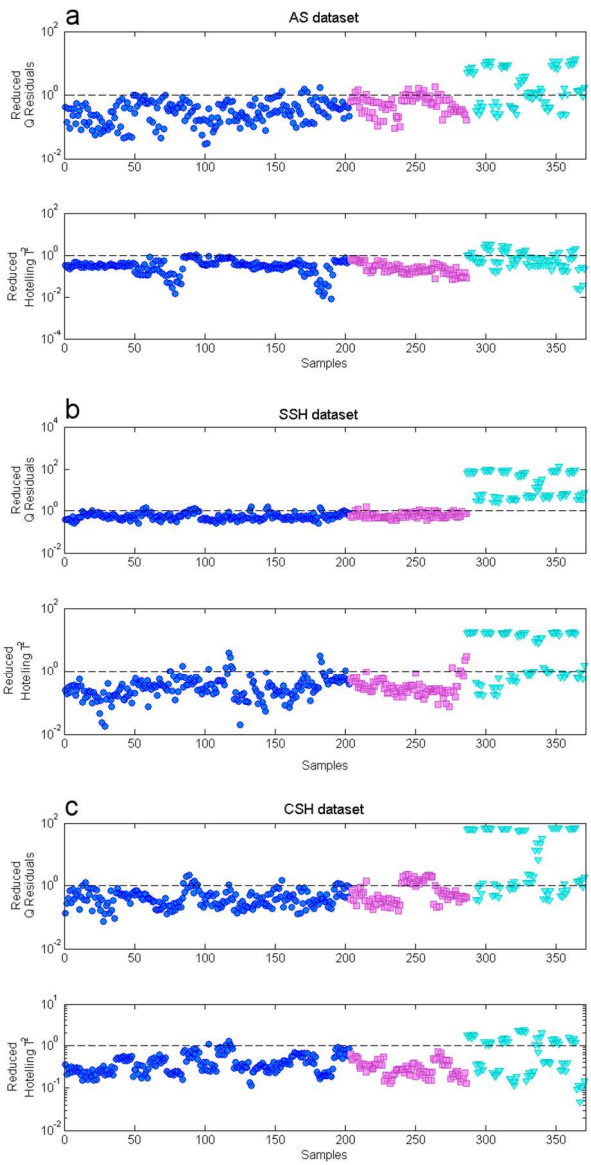

Figure 4. Q residuals and Hotelling T^2 plots of the PCA models calculated on AS (a), SSH (b) and CSH (c) datasets. Training set objects are represented with blue circles, test set objects of day-3 delivery with pink squares, and test set objects of day-4 delivery with cyan triangles. The dashed lines correspond to the 95 % confidence limit.
218x425mm (96 x 96 DPI)

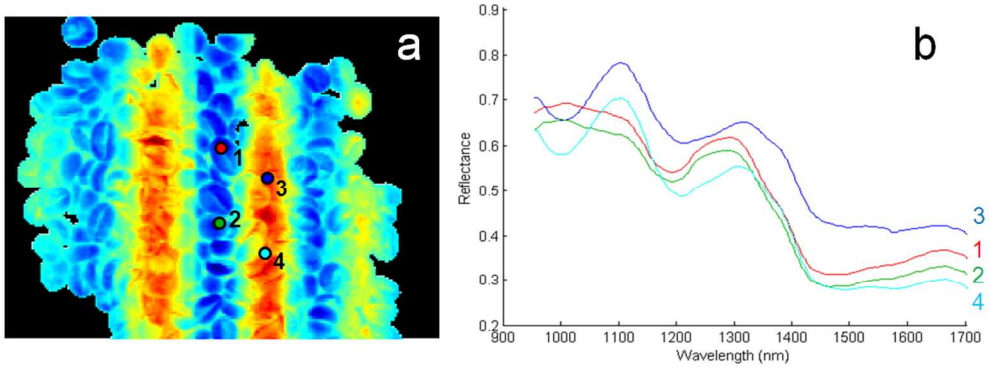

Figure 5. PC3 score image of a coffee sample acquired on day-4 presenting the striping effect (a) and reflectance spectra of different points of the same sample (b). The spectra are colored and numbered as the points to which they correspond in the score image.
419x159mm (72 x 72 DPI)

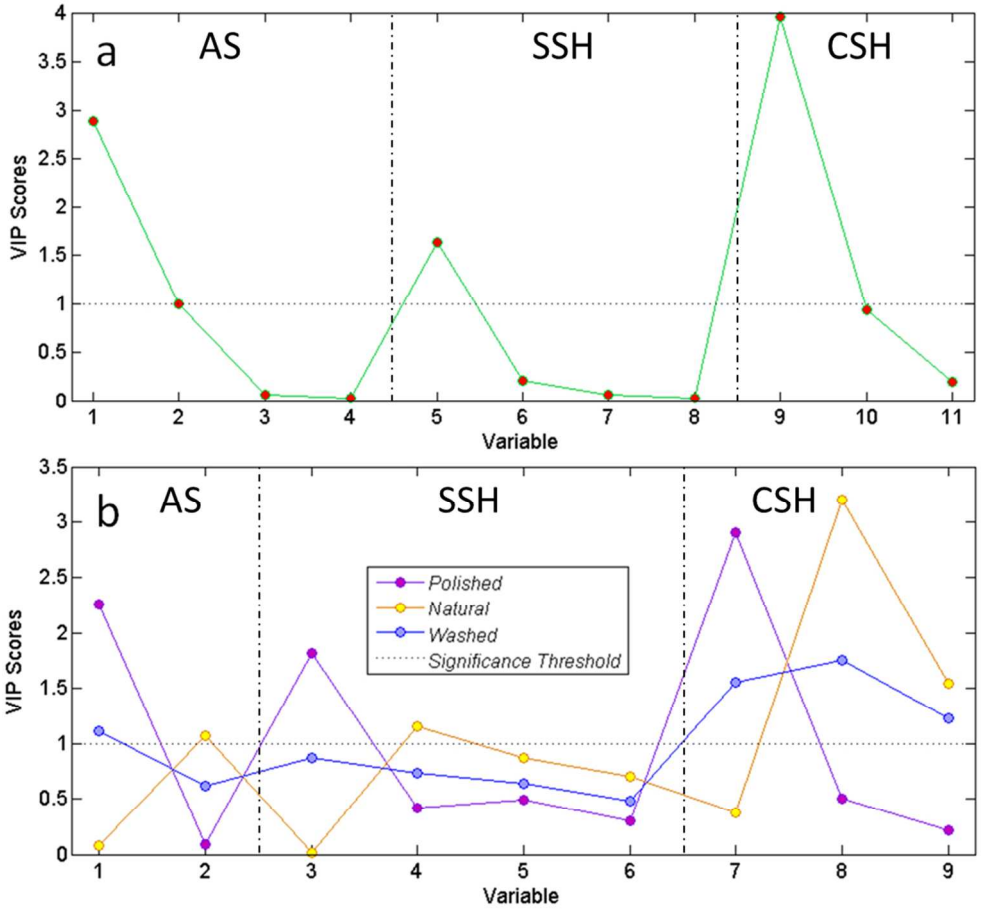

Figure 6. VIP scores of PLS-DA models calculated with block-scaled Mid-L datasets for *Arabica/Robusta* classification (a) and *Polished/Natural/Washed* classification (b).
190x175mm (150 x 150 DPI)

Day of delivery	# of delivered batches	# of acquired images	# of Arabica / Robusta	# of Natural / Washed / Polished	Data set
day 1 (April 14th)	7	84	4 / 3	3 / 3 / 1	Training set
day 2 (June 6th)	10	119	4 / 6	3 / 5 / 2	Training set
day-3 (July 3rd)	7	83	3 / 4	3 / 3 / 1	Test set day-3
day-4 (August 28th)	7	84	3 / 4	3 / 3 / 1	Test set day-4

Table 1. Description of the set of green coffee batches.

	Dataset	AS		SSH		CSH		Low-L		Mid-L		Mid-L	
	Preprocessing	SNV + mean centering		mean centering		mean centering		block-scaling		block-scaling		autoscaling	
	# of LVs	4		4		3		4		2		1	
		TP	TN	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
R	Calibration	100,0	100,0	98,1	99,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Cross-validation	99,1	100,0	90,7	95,8	88,8	100,0	88,8	100,0	88,8	100,0	100,0	100,0
	Prediction	100,0	83,3	97,9	83,3	100,0	100,0	100,0	97,2	100,0	100,0	100,0	88,9
A	Calibration	100,0	100,0	99,0	98,1	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Cross-validation	100,0	99,1	95,8	90,7	100,0	88,8	100,0	88,8	100,0	88,8	100,0	100,0
	Prediction	83,3	100,0	83,3	97,9	100,0	100,0	97,2	100,0	100,0	100,0	88,9	100,0
NER	Calibration	100,0		98,5		100,0		100,0		100,0		100,0	
	Cross-validation	99,5		93,1		94,1		94,1		94,1		100,0	
	Prediction	92,8		91,6		100,0		98,8		100,0		95,2	

Table 2. Results of PLS-DA models for *Arabica/Robusta* classification

	Dataset	AS		SSH		CSH		Low-L		Mid-L		Mid-L	
	Preprocessing	SNV + mean centering		mean centering		mean centering		block-scaling		block-scaling		autoscaling	
	# of LVs	2		4		3		4		4		3	
		TP	TN	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
P	Calibration	100,0	94,6	100,0	98,2	100,0	100,0	100,0	99,4	100,0	100,0	100,0	98,8
	Cross-validation	100,0	92,2	100,0	96,4	100,0	100,0	100,0	95,8	100,0	100,0	100,0	98,8
	Prediction	91,7	98,6	91,7	88,7	91,7	91,5	91,7	93,0	91,7	100,0	91,7	93,0
N	Calibration	66,2	90,9	74,6	84,1	100,0	87,1	97,2	94,7	100,0	98,5	97,2	99,2
	Cross-validation	66,2	90,2	63,4	75,8	85,9	81,8	76,1	77,3	100,0	86,4	95,8	96,2
	Prediction	63,9	74,5	66,7	85,1	69,4	100,0	61,1	97,9	80,6	100,0	72,2	100,0
W	Calibration	78,1	77,6	77,1	85,0	82,3	100,0	92,7	99,1	97,9	100,0	99,0	100,0
	Cross-validation	72,9	77,6	62,5	77,6	75,0	90,7	64,6	86,9	81,3	100,0	94,8	99,1
	Prediction	68,6	75,0	65,7	79,2	97,1	85,4	94,3	77,1	100,0	83,3	100,0	87,5
NER	Calibration	77,8		80,3		91,6		95,6		99,0		98,5	
	Cross-validation	75,4		69,5		83,3		74,9		91,1		96,1	
	Prediction	69,9		69,9		84,3		79,5		90,4		86,7	

Table 3. Results of PLS-DA models for *Polished/Natural/Washed* classification

ONLINE RESOURCE: SUPPLEMENTARY FIGURES AND TABLES

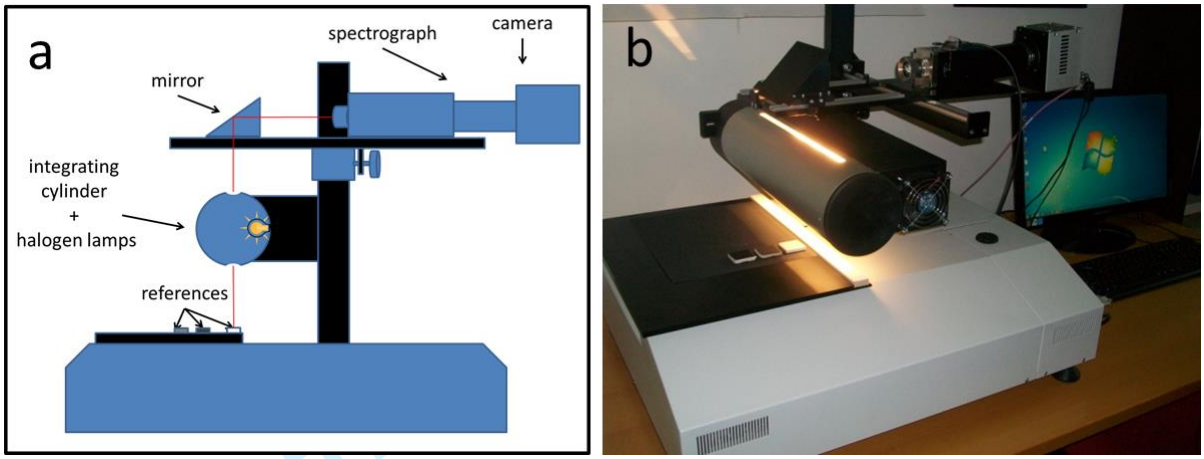

Figure S1. Schematic representation (a) and picture (b) of the hyperspectral system used for image acquisition.

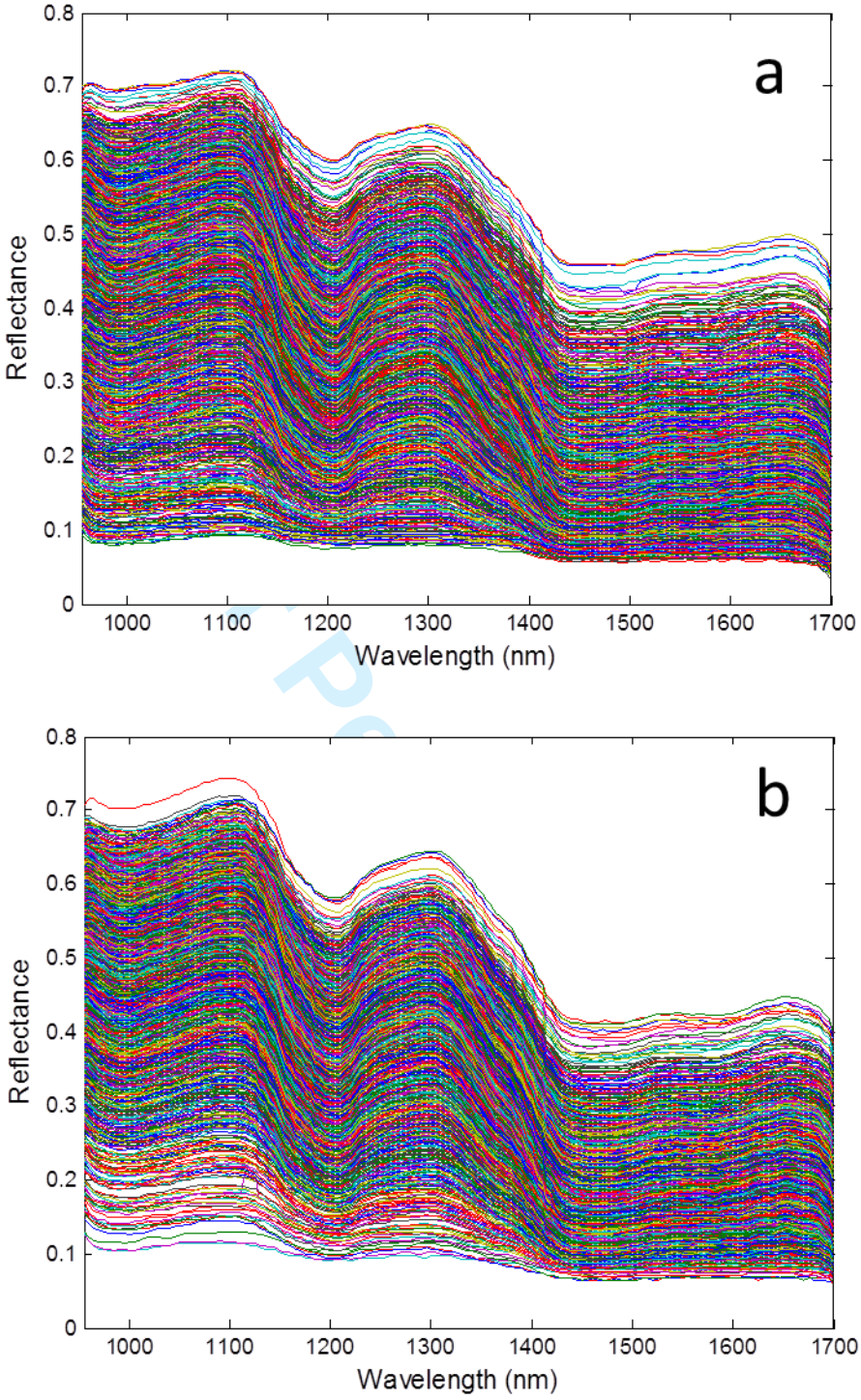

Figure S2. Original spectra retained after background elimination of one image of *Arabica* coffee (a) and one image of *Robusta* coffee (b).

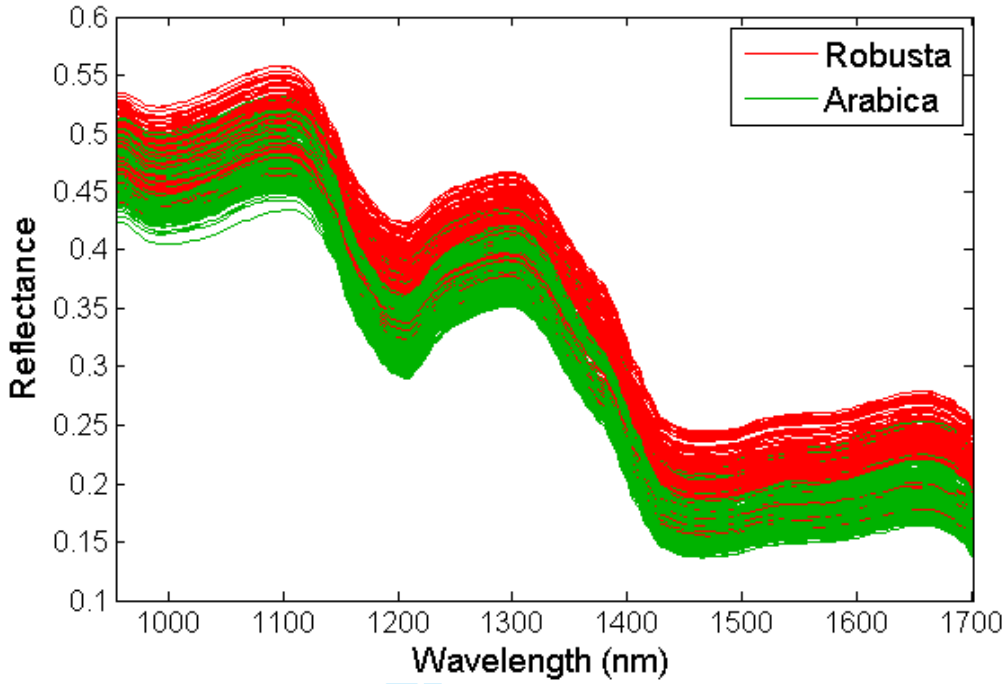

Figure S3. Average spectra of each hyperspectral image contained in AS dataset

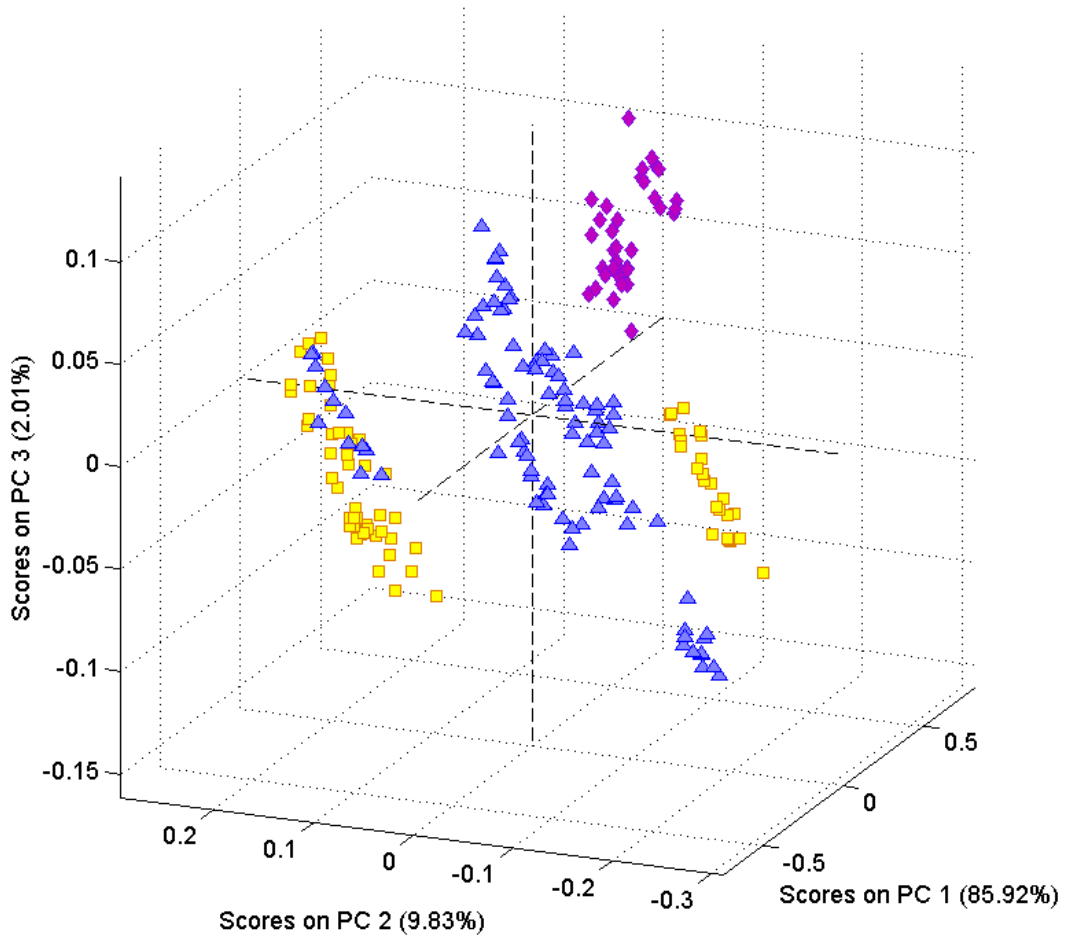

Figure S4. Score plot of PC1, PC2 and PC3, calculated for AS dataset. Purple diamonds for *Polished* coffee, blue triangles for *Washed* coffee and yellow squares for *Natural* coffee.

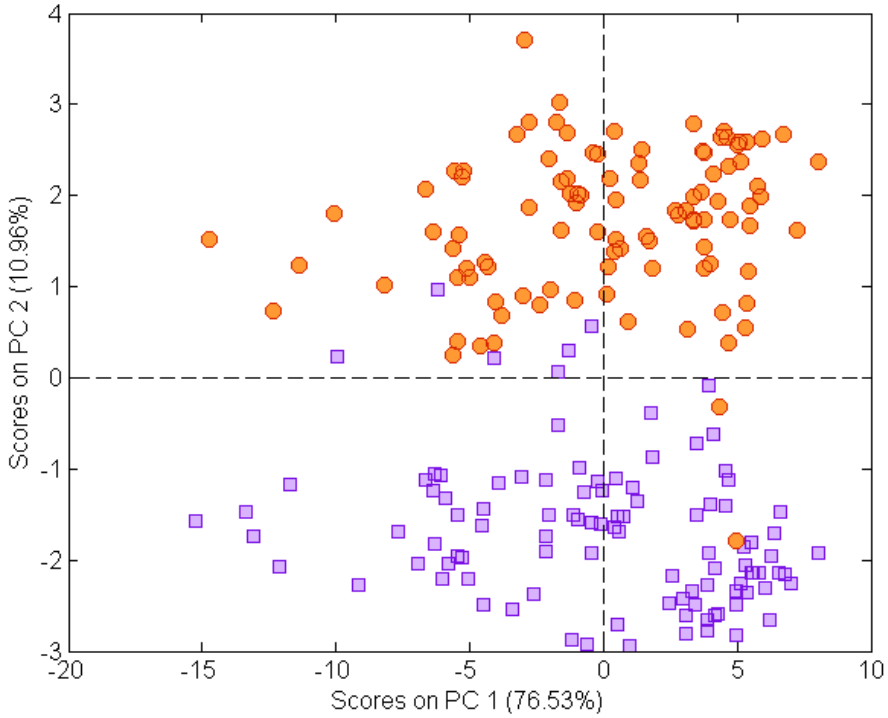

Figure S5. PC1 vs. PC2 scores plot obtained for SSH dataset. Orange circles for levelled coffee beans, purple squares for piled beans.

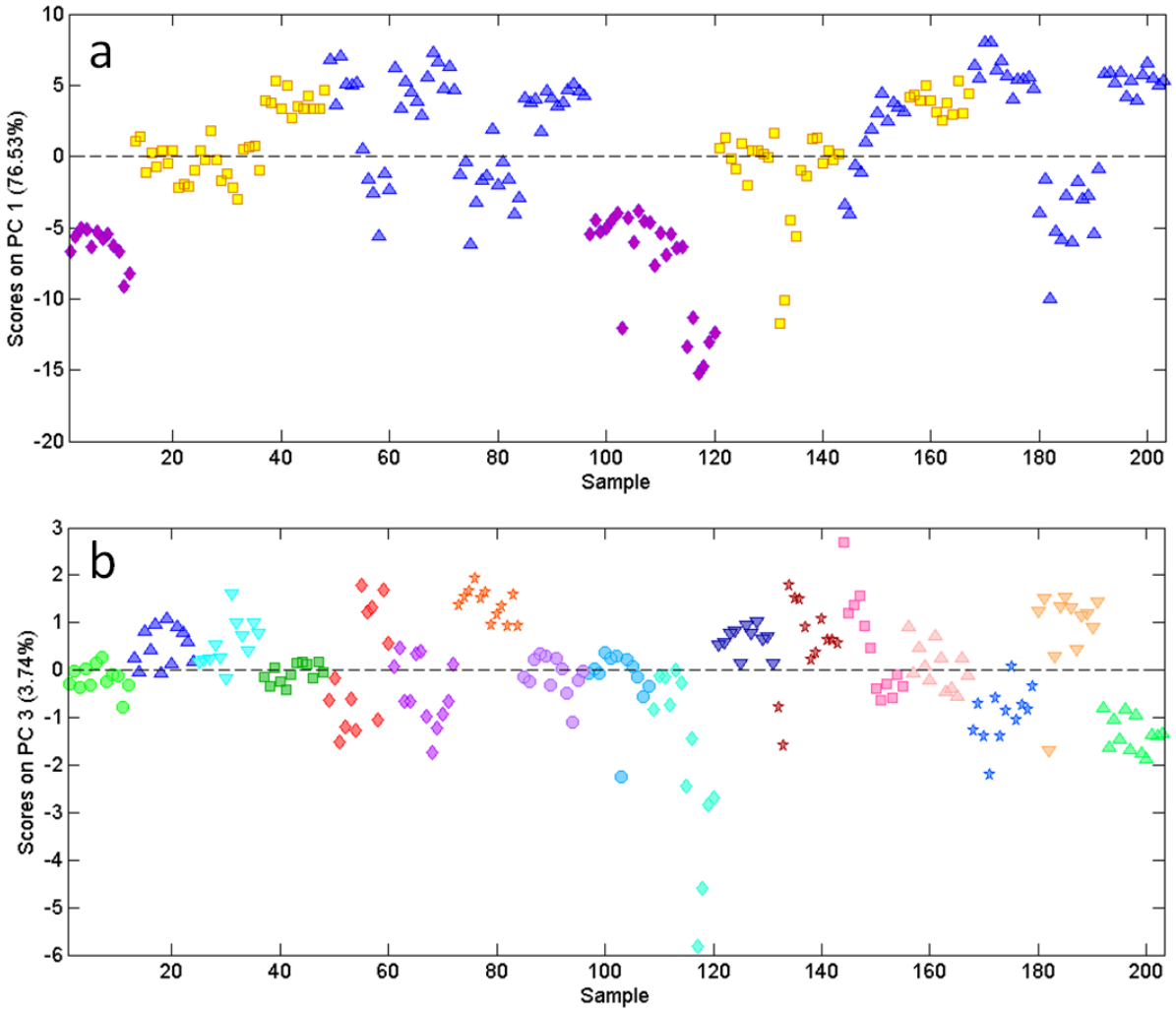

Figure S6. In (a) PC1 score plot of SSH dataset: purple diamonds for *Polished* coffee, blue triangles for *Washed* coffee and yellow squares for Natural coffee; in (b) PC3 score plot of SSH dataset: the different symbols are referred to different production batches.

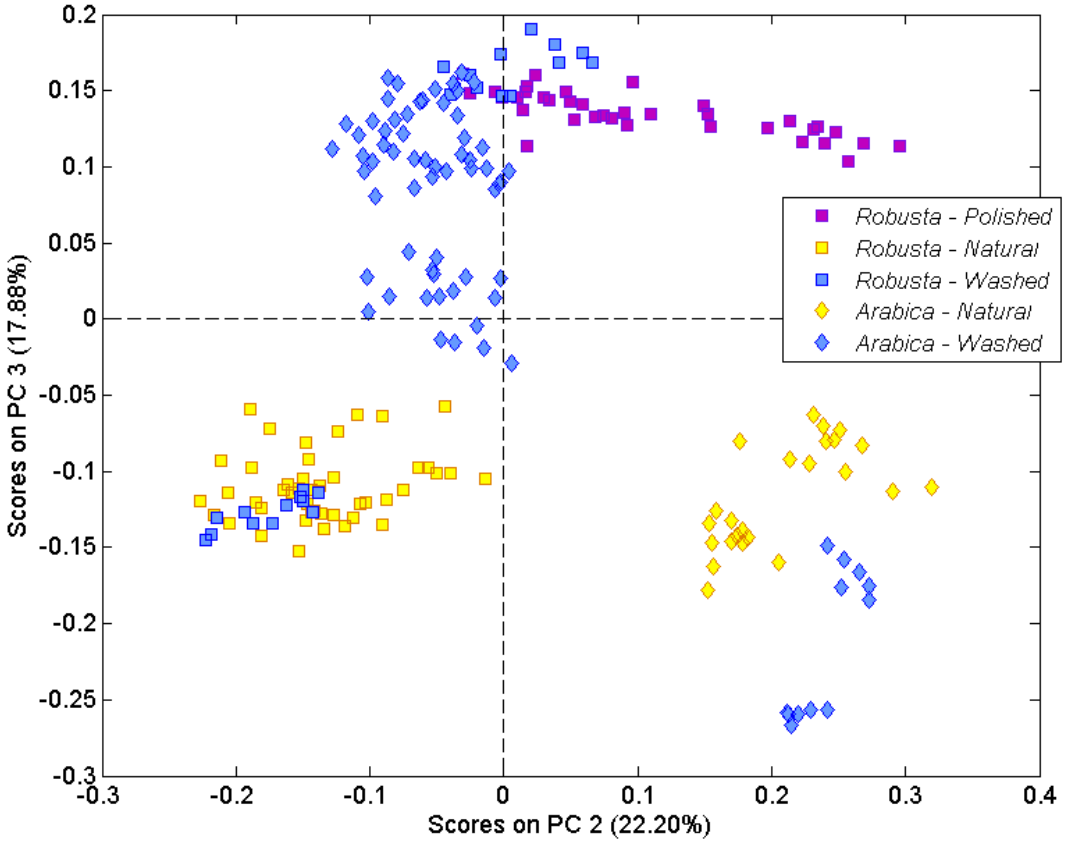

Figure S7. PC2 vs. PC3 score plot of CSH dataset.

Dataset	Preprocessing	# of LVs	Calibration				Cross-Validation				Prediction			
					Actual Class				Actual Class						Actual Class	
					R	A			R	A			R	A
AS	SNV + mean centering	4	Pred. Class	R	107	0	Pred. Class	R	106	0	Pred. Class	R	47	6
				A	0	96		A	1	96		A	0	30
SSH	mean centering	4			Actual Class				Actual Class				Actual Class	
					R	A			R	A			R	A
			Pred. Class	R	105	1	Pred. Class	R	97	4	Pred. Class	R	46	6
				A	2	95		A	10	92		A	1	30
CSH	mean centering	3			Actual Class				Actual Class				Actual Class	
					R	A			R	A			R	A
			Pred. Class	R	107	0	Pred. Class	R	95	0	Pred. Class	R	47	0
				A	0	96		A	12	96		A	0	36
Low-L	block-scaling	4			Actual Class				Actual Class				Actual Class	
					R	A			R	A			R	A
			Pred. Class	R	107	0	Pred. Class	R	95	0	Pred. Class	R	47	1
				A	0	96		A	12	96		A	0	35
Mid-L	block-scaling	2			Actual Class				Actual Class				Actual Class	
					R	A			R	A			R	A
			Pred. Class	R	107	0	Pred. Class	R	95	0	Pred. Class	R	47	0
				A	0	96		A	12	96		A	0	36
Mid-L	autoscaling	1			Actual Class				Actual Class				Actual Class	
					R	A			R	A			R	A
			Pred. Class	R	107	0	Pred. Class	R	107	0	Pred. Class	R	47	4
				A	0	96		A	0	96		A	0	32

Table S1. Confusion matrices of PLS-DA models for *Arabica/Robusta* classification.

Dataset	Preprocessing	# of LVs	Calibration				Cross-Validation				Prediction			
				Actual Class				Actual Class				Actual Class		
				P	N	W		P	N	W		P	N	W
AS	SNV + mean centering	2	Pred. Class P	36	0	9	Pred. Class P	36	0	13	Pred. Class P	11	1	0
			Pred. Class N	0	47	12	Pred. Class N	0	47	13	Pred. Class N	1	23	11
			Pred. Class W	0	24	75	Pred. Class W	0	24	70	Pred. Class W	0	12	24
SSH	mean centering	4	Pred. Class P	36	2	1	Pred. Class P	36	2	4	Pred. Class P	11	2	6
			Pred. Class N	0	53	21	Pred. Class N	0	45	32	Pred. Class N	1	24	6
			Pred. Class W	0	16	74	Pred. Class W	0	24	60	Pred. Class W	0	10	23
CSH	mean centering	3	Pred. Class P	36	0	0	Pred. Class P	36	0	0	Pred. Class P	11	5	1
			Pred. Class N	0	71	17	Pred. Class N	0	61	24	Pred. Class N	0	25	0
			Pred. Class W	0	0	79	Pred. Class W	0	10	72	Pred. Class W	1	6	34
Low-L	block-scaling	4	Pred. Class P	36	1	0	Pred. Class P	36	3	4	Pred. Class P	11	4	1
			Pred. Class N	0	69	7	Pred. Class N	0	54	30	Pred. Class N	0	22	1
			Pred. Class W	0	1	89	Pred. Class W	0	14	62	Pred. Class W	1	10	33
Mid-L	block-scaling	4	Pred. Class P	36	0	0	Pred. Class P	36	0	0	Pred. Class P	11	0	0
			Pred. Class N	0	71	2	Pred. Class N	0	71	18	Pred. Class N	0	29	0
			Pred. Class W	0	0	94	Pred. Class W	0	0	78	Pred. Class W	1	7	35
Mid-L	autoscaling	3	Pred. Class P	36	2	0	Pred. Class P	36	2	0	Pred. Class P	11	5	0
			Pred. Class N	0	69	1	Pred. Class N	0	68	5	Pred. Class N	0	26	0
			Pred. Class W	0	0	95	Pred. Class W	0	1	91	Pred. Class W	1	5	35

Table S2. Confusion matrices of PLS-DA models for *Polished/Natural/Washed* classification.