

This is the peer reviewed version of the following article:

Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee / Calvini, Rosalba; Amigo, Jose Manuel; Ulrici, Alessandro. - In: ANALYTICA CHIMICA ACTA. - ISSN 0003-2670. - 967:(2017), pp. 33-41. [10.1016/j.aca.2017.03.011]

*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

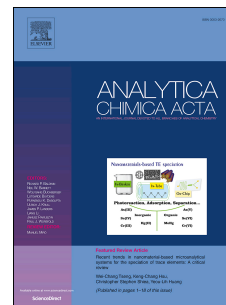
29/04/2026 14:26

(Article begins on next page)

# Accepted Manuscript

Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems:  
A filter-based simulation applied to the classification of Arabica and Robusta green  
coffee

Rosalba Calvini, Jose Manuel Amigo, Alessandro Ulrici



PII: S0003-2670(17)30289-1

DOI: [10.1016/j.aca.2017.03.011](https://doi.org/10.1016/j.aca.2017.03.011)

Reference: ACA 235112

To appear in: *Analytica Chimica Acta*

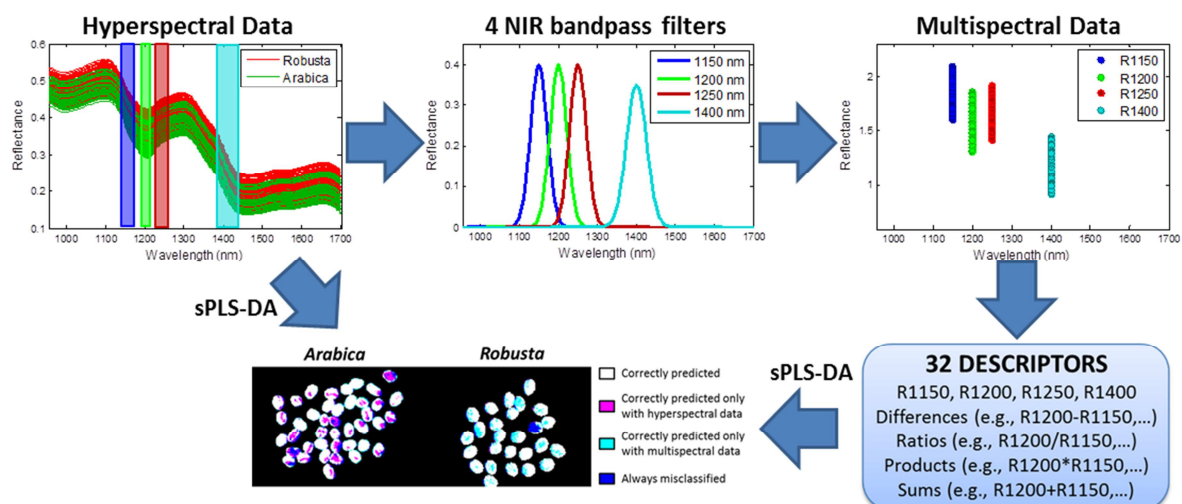
Received Date: 4 October 2016

Revised Date: 2 March 2017

Accepted Date: 6 March 2017

Please cite this article as: R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Analytica Chimica Acta* (2017), doi: 10.1016/j.aca.2017.03.011.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



1     **Transferring results from NIR-hyperspectral to NIR-multispectral**  
2     **imaging systems: a filter-based simulation applied to the classification**  
3     **of Arabica and Robusta green coffee**

4             Rosalba Calvini <sup>a,\*</sup>, Jose Manuel Amigo <sup>b</sup>, Alessandro Ulrici <sup>a</sup>

5             <sup>a</sup> *Department of Life Sciences, University of Modena and Reggio Emilia, Padiglione Besta, Via*  
6             *Amendola 2, 42122 Reggio Emilia, Italy*

7             <sup>b</sup> *Department of Food Science, Faculty of Sciences, University of Copenhagen, Rolighedsvej 30,*  
8             *DK-1958 Frederiksberg C, Denmark*

9  
10     \* Corresponding author: Rosalba Calvini, Department of Life Sciences, University of Modena and  
11     Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy. Tel: +39 0522  
12     5222048. Fax: +39 0522 522027. E-mail: [rosalba.calvini@unimore.it](mailto:rosalba.calvini@unimore.it).

13  
14     **ABSTRACT**

15     Due to the differences in terms of both price and quality, the availability of effective  
16     instrumentation to discriminate between Arabica and Robusta coffee is extremely important. To this  
17     aim, the use of multispectral imaging systems could provide reliable and accurate real-time  
18     monitoring at relatively low costs. However, in practice the implementation of multispectral  
19     imaging systems is not straightforward: the present work investigates this issue, starting from the  
20     outcome of variable selection performed using a hyperspectral system. Multispectral data were  
21     simulated considering four commercially available filters matching the selected spectral regions,  
22     and used to calculate multivariate classification models with Partial Least Squares-Discriminant  
23     Analysis (PLS-DA) and sparse PLS-DA. Proper strategies for the definition of the training set and

1 the selection of the most effective combinations of spectral channels led to satisfactory  
2 classification performances (100% classification efficiency in prediction of the test set).

3

4 **KEYWORDS**

5 Green coffee; Multivariate classification; Hyperspectral imaging; Multispectral imaging; Sparse  
6 methods.

7

ACCEPTED MANUSCRIPT

## 1 1. INTRODUCTION

2 Hyperspectral imaging (HSI) systems have showed a great potential for their application in food  
3 control processes, being a fast and non-destructive technique able to handle different issues related  
4 to quality evaluation and safety inspection of agricultural and food products [1 – 3]. The main  
5 advantage of HSI is the possibility of coupling classical point-wise spectroscopy with imaging  
6 techniques, in order to obtain simultaneously spatial and spectral information from a sample. In this  
7 manner, a huge amount of spectral information is achieved from the surface of a sample in a  
8 relatively short time and it is possible to provide a reliable and accurate real-time monitoring at  
9 different stages of the food processing chain [4].

10 Despite the advantages of this technique, two main drawbacks currently limit the direct  
11 implementation of HSI into real-time systems for food control: the high costs of HSI systems and  
12 the extremely large amount of data that are acquired in short times, implying high computational  
13 loads which complicate the development of efficient and fast applications [5, 6]. For these reasons,  
14 most of the food related HSI research works have been directed towards the identification of  
15 wavelengths relevant to the problem at hand, for further development of multispectral imaging  
16 systems suitable for on line or portable devices [3, 4, 7]. The advantages of multispectral imaging  
17 systems over hyperspectral ones include the faster acquisition times and the lower costs of  
18 hyperspectral cameras [8]. Furthermore, multispectral technology is easier to implement for on-line  
19 applications, due to the higher resistance and stability of the optical components.

20 Starting from variable selection performed on hyperspectral image data, a multispectral imaging  
21 system can be implemented, at least in principle, by using the selected spectral regions [9 – 12].  
22 However, adapting the outcomes of variable selection performed on hyperspectral data to a filter-  
23 based imaging system is not straightforward, and it is not easy to maintain acceptable performances  
24 after transferring results from a hyperspectral imaging system to a multispectral imaging one [10].  
25 In fact, position and width of actually available filters generally do not perfectly match the selected

1 spectral regions. Moreover, in multispectral systems only a single “average” intensity value can be  
2 measured for each selected spectral region, which implies that the useful information related to  
3 spectral shape within the selected intervals is lost. On the other hand, the limited number of  
4 multispectral channels allows to easily expand the number of potentially useful descriptors by  
5 calculating quantities derived from the single channel intensities, i.e., by introducing nonlinear and  
6 interaction terms derived from the intensity values in order to improve the results [13]. In this  
7 manner, it is possible to evaluate linear and non-linear relationships between the different channels  
8 and thus to emphasize small variations in the spectral signature which could be useful for the  
9 problem at hand [14].

10 In this context, the present work is aimed at showing the feasibility of implementing a simulated  
11 multispectral filter-based classification model for the discrimination of green coffee samples  
12 belonging to Arabica (*Coffea arabica*) and Robusta (*Coffea canephora*) coffee species. Arabica and  
13 Robusta coffee differ each other from chemical and organoleptic properties. Chemical analysis  
14 based on chromatography showed that Arabica and Robusta coffee have differences in their content  
15 in caffeine, chlorogenic acid, trigonelline, sterols and amino acids [15, 16]. Furthermore, Arabica  
16 coffee is considered of higher quality because of its better taste and aroma and, therefore, it is  
17 generally preferred by the consumers. On the other hand, Arabica coffee can be two to ten times  
18 more expensive than Robusta coffee, which is less appreciated and mainly used as filler in coffee  
19 blends or in instant coffee production [17, 18]. Due to the significant differences in terms of both  
20 price and quality, it is therefore important to correctly discriminate between the two coffee coffee  
21 species in order to prevent the adulteration of high quality Arabica coffee with cheaper and lower  
22 quality Robusta coffee [19 - 22]. Therefore, the correct classification of green coffee beans of the  
23 two species could allow to identify possible adulterations or mislabelling at an early stage of the  
24 processing chain.

25 Many different analytical techniques have been investigated in order to discriminate Arabica and  
26 Robusta coffee species, including  $^1\text{H}$ -Nuclear Magnetic Resonance (NMR) [23],  $^{13}\text{C}$ -NMR [24], 2-

1 D electrophoresis [25], electronic nose and electronic tongue [26] among others. The use of so  
2 many different analytical methods suggest that a reliable discrimination between Arabica and  
3 Robusta species is a crucial aspect for coffee industry.

4 However, the above mentioned methods are not suitable for fast characterization of large amounts  
5 of products, e.g., by rapid in-line or on-line monitoring [19]; conversely, spectroscopic methods  
6 represent a fast and non-destructive alternative to other more complex analytical techniques for  
7 facing food authentication problems, being at the same time simple, fast, non-destructive and  
8 reliable [27]. In particular, classical point-wise NIR spectroscopy has been widely used to  
9 discriminate Arabica and Robusta species, both on green coffee [28, 29] and on roasted coffee [30,  
10 31]. More recently, some research works have been also reported where hyperspectral imaging  
11 (HSI) is used to characterize these coffee species [32, 33].

12 The main aim of the present study is to investigate the issues related to the implementation of a  
13 multispectral imaging system starting from the outcome of a variable selection/classification model  
14 calculated on hyperspectral data [7]. In particular, a multispectral detection system was simulated  
15 by considering four commercially available filters, chosen as those showing the best match with the  
16 selected spectral regions. Then, Partial Least Squares Discriminant Analysis (PLS-DA) [6, 34, 35]  
17 classification models were built considering as descriptors both the four channels alone and the four  
18 channels together with their squared values and sums, differences, products and ratios between  
19 couples of channels. Moreover, variable selection by Sparse Partial Least Squares Discriminant  
20 Analysis (sPLS-DA) [7, 36] was also employed in order to identify the most relevant descriptors  
21 and to further increase the performances of the classification models.

22

## 1 **2. MATERIALS AND METHODS**

### 2 **2.1 Coffee samples**

3 The green coffee samples of Arabica and Robusta species considered in this study came from  
4 different geographical areas and were provided by a local roasting company during a period of 6  
5 months. In particular, each sample belonged to a different batch and the samples were subjected to  
6 different processing methods to separate the seed from the fruit. Despite the different properties of  
7 the samples, we focused on the discrimination between Arabica and Robusta coffee species,  
8 regardless of processing method or of geographical origin.

9 On the whole, 33 batches were considered in the industrial plant: 18 of Robusta and 15 of Arabica.  
10 From each batch about 500 g of beans, sampled in order to be as representative as possible of the  
11 corresponding batch, were collected and stored in a sealed package. From each package, three  
12 aliquots of 70 g of randomly selected beans were taken, and for each aliquot two repeated images  
13 were acquired, shuffling the beans between the two acquisitions. The same procedure was repeated  
14 in a different day in order to check the day-to-day variability. All the batches were acquired in  
15 random order and the packages were sealed again and stored at room temperature between the  
16 different acquisition days. Therefore, for each batch 12 hyperspectral images (= 2 measurement  
17 sessions  $\times$  3 aliquots  $\times$  2 repeated acquisitions) were acquired, obtaining a dataset composed by 396  
18 hyperspectral images (33 batches  $\times$  12 images).

19

### 20 **2.2 Image acquisition**

21 The hyperspectral images were acquired using a desktop NIR Spectral Scanner (DV Optic)  
22 embedding a reflectance imaging based spectrometer Specim N17E, coupled to a Xenics XEVA  
23 2608 camera (320  $\times$  256 pixels) and working in the 955-1700 nm spectral range with a spectral  
24 resolution of 5 nm, with a total of 150 spectral channels. The images were acquired using a black  
25 silicon carbide sandpaper sheet as background, which is characterized by a very low and constant

1 reflectance spectrum [37], and in addition a 99% white ceramic tile reflectance standard and two  
2 ceramic tiles with intermediate reflectance values were included in the area of the images.

3 The raw data were converted into reflectance values using an instrumental calibration based on the  
4 high reflectance standard reference and on dark current [38]. Furthermore, in order to reduce the  
5 variability among images over time, an additional internal calibration was performed [6, 39], based  
6 on the average reflectance values of the reflectance standard, of the two ceramic tiles and of the  
7 black silicon carbide sandpaper.

8 Before further analysis, the pixels related to the black sandpaper background were removed from  
9 each image using the following thresholding procedure: a preliminary evaluation of some sample  
10 images allowed to identify the most discriminant wavelength by maximizing the Fisher ratio  
11 between background spectra and sample spectra. In this manner, at 1050 nm, all the pixels below  
12 the threshold value of 0.1 reflectance units were identified as background and removed.

## 14 **2.3 Data analysis**

### 15 *2.3.1 Data arrangement*

16 The acquired hyperspectral images were used to create different hyperspectral datasets, which in  
17 turn were converted into the corresponding multispectral datasets and then used for calculation and  
18 validation of the classification models.

19 In particular, the 33 coffee batches were randomly split in 24 training batches (corresponding to 288  
20 training set images), including 11 Arabica and 13 Robusta coffee batches, and in 9 test batches (108  
21 images), including 4 Arabica and 5 Robusta coffee batches.

22 Two different strategies were then considered for the definition of the training set:

- 23 - Average Spectra (AS): the average spectrum was calculated for each training set image,  
24 obtaining an AS training set including 288 spectra;
- 25 - Random spectra (RS): 50 spectra were randomly selected from each training set image,  
26 obtaining a RS training set including 14400 spectra (= 288 images  $\times$  50 spectra).

1 Since each hyperspectral image is composed by tens of thousands of single-pixel spectra, AS  
2 training set and RS training set represent two different approaches to build a reduced (but still  
3 representative) training set from the huge amount of data contained in the original images.

4 On one hand, computing the average spectra from each hyperspectral image allows to drastically  
5 reduce the number of objects, under the assumption that an average spectrum is representative of  
6 the image as a whole. On the other hand, averaging all the pixel-spectra contained in each image  
7 implies losing the information about the spatial variability within each image. Therefore, keeping a  
8 relatively limited number of randomly selected spectra for each image was considered as a valid  
9 compromise between reducing data size and (at least partly) maintaining spatial variability related  
10 information.

11 The classification performances of the two approaches were evaluated in prediction both at the  
12 image-level, i.e., considering each image on the whole, and at the pixel-level, i.e., evaluating the  
13 class assignment of each single image pixel. In particular, an external test set consisting of the  
14 average spectra of each image of the test samples was used to validate the classification models at  
15 the image-level. Moreover, in order to validate the classification models also at the pixel-level and  
16 to visually evaluate the classification performances, two different test images (test image 1 and test  
17 image 2) with different arrangement of the beans were considered. The test images were obtained  
18 by merging together one image of Arabica coffee and one of Robusta coffee taken from the test  
19 samples. In this manner, since each test image contains one image for each class, it was possible to  
20 obtain a quantitative evaluation of the predictive ability of the models at the pixel-level.

21 The hyperspectral datasets (AS training set, RS training set, test set and the two test images)  
22 mentioned in this section were then converted into multispectral data following the procedure  
23 described in below.

24

### 1 2.3.2 From HSI selected spectral regions to multispectral data

2 Number and position of the channels to be considered for the multispectral system were defined on  
3 the basis of feature selection made by applying sparse classification methods to the AS training set  
4 described in the previous section, as reported in Calvini et al., 2015 [7]. Briefly, sparse methods  
5 allow performing variable selection by forcing the model coefficients related to noisy or  
6 uninformative variables to be equal to zero, in a way that it is possible to calculate a classification  
7 model and to perform variable selection in a one-step procedure [40]. For the classification of  
8 Arabica and Robusta green coffee two different sparse-based algorithms, i.e., sparse Principal  
9 Component Analysis [41] coupled with k-Nearest Neighbour [42] (sPCA+kNN) and sparse Partial  
10 Least Squares Discriminant Analysis (sPLS-DA) [36], were applied to the average spectra of the  
11 hyperspectral images. The relevance of the selected wavelengths, which are reported in Figure 1.a,  
12 was confirmed by the fact that both sparse classification methods converged to the selection of the  
13 same narrow spectral regions. The evaluation of these regions showed that they essentially reflect  
14 chemical composition differences between the two coffee varieties, rather than physical effects. In  
15 particular, the selected spectral regions were related to the C-H aromatic second overtone (1143  
16 nm), to the C-H aliphatic second overtone (1195-1225 nm), and to the O-H first overtone of  
17 aliphatic (1410 nm) and aromatic alcohol (1420 nm) [43]. In order to better evaluate the spectral  
18 differences between Arabica and Robusta green coffee, Figure S-1 reports two sample spectra  
19 belonging to the two coffee species and the corresponding difference spectrum.

20 Starting from the outcome of variable selection performed on hyperspectral data, the commercial  
21 filters showing the best match with the selected spectral regions were then identified [44]; in  
22 particular, four bandpass filters were selected, whose center wavelength (*CWL*), filter width at half  
23 maximum (*FWHM*) and peak transmission (*PT*) values listed below:

- 24 - Filter 1: *CWL*=1150 nm, *FWHM*=10 nm, *PT*=40%
- 25 - Filter 2: *CWL*=1200 nm, *FWHM*=10 nm, *PT*=40%
- 26 - Filter 3: *CWL*=1250 nm, *FWHM*=10 nm, *PT*=40%

1 - Filter 4: CWL=1400 nm, FWHM=12 nm, PT=35%.

2 In order to mimic the output of a filter-based multispectral system, each reflectance spectrum of the  
3 hyperspectral datasets (AS training set, RS training set, test set and the two test images) was then  
4 used to estimate the reflectance values that would be obtained by using the considered filters.

5 To this purpose, the Gaussian-shaped transmission profile of each filter was calculated from the  
6 corresponding filter properties as:

$$7 \quad Z(\lambda) = PT \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\lambda-CWL)^2}{2\sigma^2}} \quad (1)$$

8 where  $Z(\lambda)$  is the Gaussian transmission profile of the filter as a function of the wavelength  $\lambda$ , and  $\sigma$   
9 is the standard deviation of the Gaussian distribution, which is related to  $FWHM$  according to the  
10 equation [45]:

$$11 \quad \sigma = \frac{FWHM}{2\sqrt{2\ln 2}} \quad (2)$$

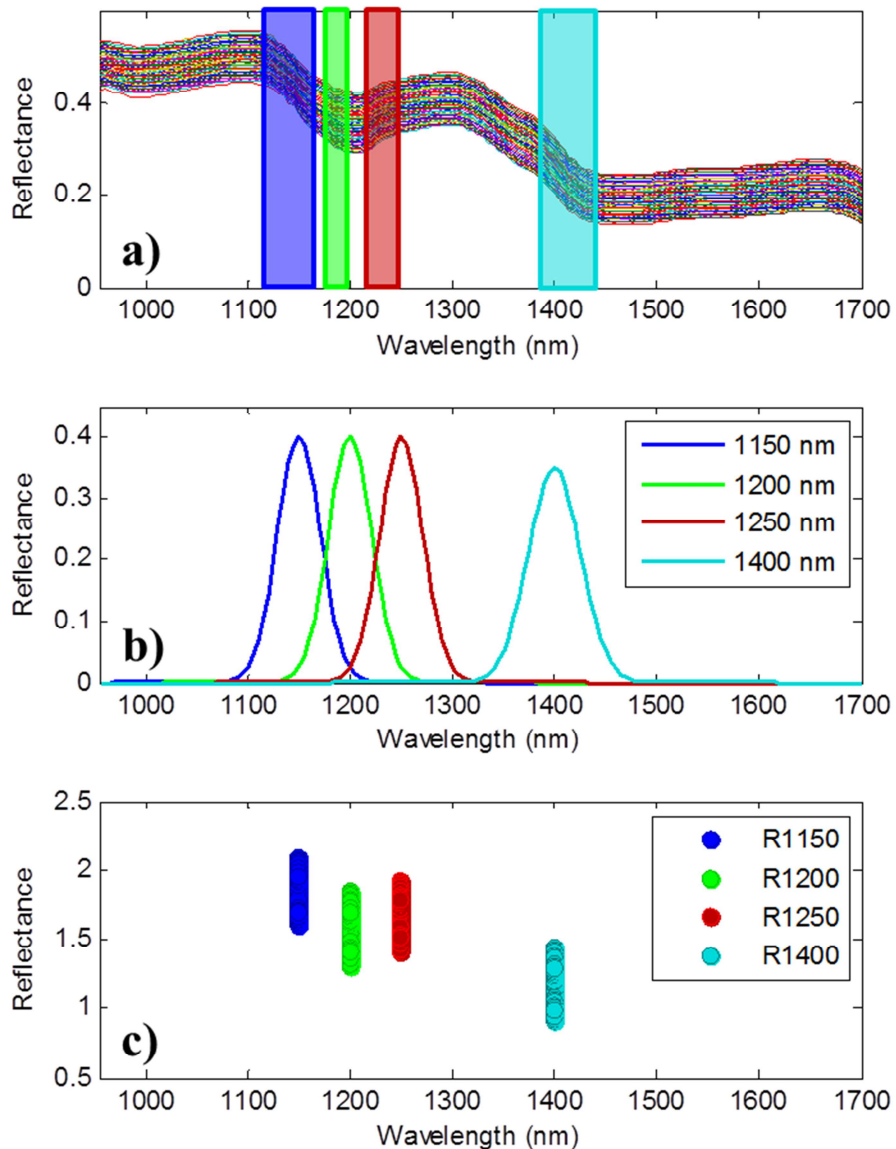
12 For each filter, the corresponding reflectance value of each object  $i$  belonging to the hyperspectral  
13 datasets was then calculated according to the equation:

$$14 \quad R_i = \sum_{\lambda_{min}}^{\lambda_{max}} Z(\lambda) S_i(\lambda) \quad (3)$$

15 where  $S_i(\lambda)$  is the HSI spectrum of object  $i$ , while  $\lambda_{min}$  and  $\lambda_{max}$  are the extreme values of the  
16 spectral range acquired with the hyperspectral camera (in our case,  $\lambda_{min} = 955$  nm and  $\lambda_{max} = 1700$   
17 nm).

18 This procedure is schematically represented in Figure 1, where Figure 1.a reports the spectra of AS  
19 training set together with the selected spectral regions, Figure 1.b shows the Gaussian profiles  
20 calculated by equation (1), and Figure 1.c reports the 4 discrete reflectance values calculated by  
21 equation (3) for each spectrum of the AS training set.

22



**Figure 1.** Average spectra calculated from each image of the training samples with selected regions highlighted (a), Gaussian profiles of the 4 considered filters (b) and resulting reflectance values obtained from the average training spectra (c).

1  
 2  
 3  
 4  
 5  
 6 Therefore, from the original hyperspectral datasets containing 150 spectral variables, the  
 7 corresponding multispectral datasets containing 4 variables were obtained, where each variable is  
 8 the reflectance value of a given multispectral channel. As an example, Figure S-2 (Supplementary  
 9 material) shows the pseudocolor images obtained from the four considered channels for both test  
 10 image 1 and test image 2.

11

1 2.3.3 Filter-based simulations

2 Starting from the multispectral datasets composed of 4 variables, three different simulations were  
3 performed:

- 4 • In *Filter Simulation 1* only the 4 reflectance values of the filters were considered for PLS-  
5 DA classification.
- 6 • In *Filter Simulation 2*, since the low number of multispectral channels allows expanding the  
7 number of potentially useful descriptors by calculating quantities derived from the outputs  
8 of the different channels, PLS-DA classification models were calculated including also  
9 additional descriptors derived from the four reflectance values. In particular, the squared  
10 reflectance value of each channel and differences, ratios, products and sums between  
11 couples of channels were calculated, obtaining datasets with the 32 descriptors listed in  
12 Table 1.
- 13 • In *Filter Simulation 3*, feature selection by sPLS-DA was performed in order to identify the  
14 most relevant descriptors among the 32 variables listed in Table 1.

15 For each of the three simulations, two classification models were calculated, one using the AS  
16 training set and one using the RS training set.

17

Single Filters	Differences	Ratios	Products	Sums
R1150	R1200-R1150	R1150/R1200	R1150*R1200	R1150+R1200
R1200	R1250-R1200	R1150/R1250	R1150*R1250	R1150+R1250
R1250	R1400-R1250	R1150/R1400	R1150*R1400	R1150+R1400
R1400	R1400-R1200	R1200/R1250	R1200*R1250	R1200+R1250
R1150 <sup>2</sup>	R1400-R1150	R1200/R1400	R1200*R1400	R1200+R1400
R1200 <sup>2</sup>	R1250-R1150	R1250/R1400	R1250*R1400	R1250+R1400
R1250 <sup>2</sup>				
R1400 <sup>2</sup>				

18 **Table 1.** List of the 32 descriptors considered.

19

#### 1 2.3.4 PLS-DA and sPLS-DA

2 PLS-DA [34] is a classification method based on the application of PLS regression to classification  
3 purposes. In PLS-DA the  $\mathbf{Y}$  matrix is composed by as many columns as the number of existing  
4 classes, where each column is a binary class vector, reporting the 1 value if the corresponding row  
5 (object) belongs to the class, and 0 otherwise. The matrix of descriptors ( $\mathbf{X}$ ) is therefore regressed  
6 against the  $\mathbf{Y}$  binary matrix, and the outcome of the PLS model is a matrix containing the regression  
7 coefficients which are used to predict new samples. The  $\mathbf{Y}$  predictions for each object with respect  
8 to each class are continuous values, therefore a threshold based on Bayesian statistics [46] has to be  
9 set separately for each class model (i.e. for the estimate of each class vector), so that objects whose  
10 predicted value for a given class is higher than the threshold are attributed to that class, while  
11 objects leading to predictions lower than the threshold are not assigned to the class.

12 Sparse PLS-DA [7, 36] is an extension of PLS-DA in which a penalty function is applied to the  
13 model parameters in order to constrain some coefficients to be equal to zero, i.e., to induce sparsity  
14 on the model coefficients. In particular, sparsity is induced on the PLS loadings, and consequently  
15 on the regression coefficients used to predict unknown samples. Therefore, thanks to the sPLS-DA  
16 approach it is possible to perform both classification and variable selection in a one-step procedure,  
17 by forcing to zero the coefficients of noisy or uninformative variables.

18 In order to select the best sPLS-DA classification models, in addition to the number of latent  
19 variables (LVs) like for PLS-DA, also the number of variables to select for each component needs  
20 to be tuned. Therefore, in Filter Simulation 3, different sPLS-DA models were constructed using  
21 both AS training set and RS training set, and testing all the combinations between a number of LVs  
22 ranging from 1 to 5 and a number of variables to select for each component ranging from 2 to 32.  
23 For each training set, the best model was selected by keeping the classification error in cross-  
24 validation as low as possible, while retaining at the same time the lowest possible number of  
25 variables and of LVs.

1 Both PLS-DA and sPLS-DA were calculated on autoscaled variables since the considered  
2 descriptors have different scales. The optimization of the classification models was performed using  
3 contiguous blocks cross-validation with 4 deletion groups, where each block contained the average  
4 spectra (for AS training set) or randomly selected spectra (for RS training set) of all the replicated  
5 and repeated images belonging to 6 batches (i.e., setting aside whole batches in the different  
6 deletion groups), as shown in Figure S-3 of Supplementary material.

7 For all the simulations, the classification performances were defined using efficiency (*EFF*), which  
8 is the geometric mean between sensitivity (*SENS*) and specificity (*SPEC*), i.e.:

$$9 \quad EFF = \sqrt{SENS \times SPEC} \quad (4)$$

10 where sensitivity is the percentage of objects of each class correctly retained by the class model and  
11 specificity is the percentage of objects of the other classes correctly rejected by the class model.

12 Data analysis was performed using PLS\_Toolbox (v. 7.5, Eigenvector Research Inc., USA) for  
13 PLS-DA, while sPLS-DA was computed with *ad-hoc* routines kindly provided by Dr. Ewa  
14 Szymanska and written in Matlab language (ver. 7.12, The Mathworks Inc., USA); further details  
15 can be found in Szymanska et al., 2015 [47]. The data were analyzed using a personal computer  
16 running with Windows 8.1-64 bit and equipped with an Intel Core® i7-3632QM CPU @ 2.20GHz  
17 processor and 6.00 GB RAM.

18

### 1 3 RESULTS AND DISCUSSION

#### 2 3.1 Filter Simulation 1

3 The results obtained from the PLS-DA classification models built both using AS training set and RS  
 4 training set and considering only the reflectance values of the 4 channels are reported in the first  
 5 two columns of Table 2. Concerning calibration and cross validation, the PLS-DA model built  
 6 using AS training set gives definitely best results than the one built with RS training set. However,  
 7 AS training set and RS training set show comparable results for the test set of average spectra and  
 8 RS training set gives better results when used to predict both test images at the pixel level. These  
 9 differences are likely due to the fact that the variability among spectra belonging to the same class  
 10 is much higher in RS training set than in AS training set. On the one hand, this led to worst results  
 11 of RS training set in calibration and cross validation, but on the other hand it turned out to be useful  
 12 for the predictions of the test images at the pixel-level.

	Filter Simulation 1		Filter Simulation 2		Filter Simulation 3	
	AS tr. set	RS tr. set	AS tr. set	RS tr. set	AS tr. set	RS tr. set
<b>N° variables</b>	4	4	32	32	4	4
<b>LVs</b>	4	4	3	3	2	2
<b><math>EFF_{CAL}</math></b>	99.1	81.3	100.0	80.3	99.6	78.4
<b><math>EFF_{CV}</math></b>	97.3	79.9	99.3	78.7	97.3	78.0
<b><math>EFF_{TEST}</math></b>	94.0	94.9	97.5	100.0	98.3	100.0
<b><math>EFF_{IMG1}</math></b>	65.1	74.0	71.1	71.0	69.5	83.9
<b><math>EFF_{IMG2}</math></b>	83.9	92.2	84.3	92.1	85.6	93.1

14 **Table 2.** Results obtained using AS training set and RS training set in filters simulation1, filter  
 15 simulation 2 and filter simulation 3; the classification efficiency values are referred to the results in  
 16 calibration ( $EFF_{CAL}$ ), cross-validation ( $EFF_{CV}$ ), prediction of the test set at the image-level  
 17 ( $EFF_{TEST}$ ), and of test image 1 and test image 2 at the pixel-level ( $EFF_{IMG1}$  and  $EFF_{IMG2}$ ,  
 18 respectively).

19

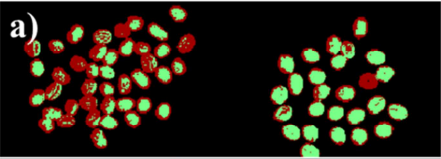
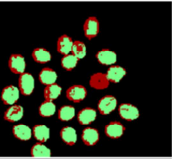
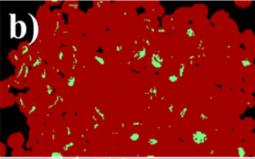
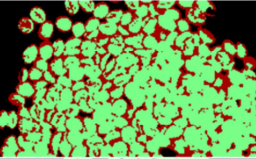
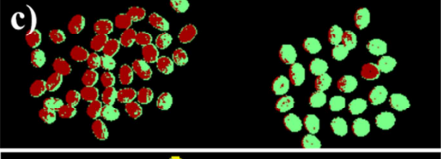
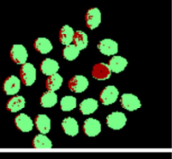
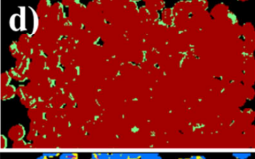
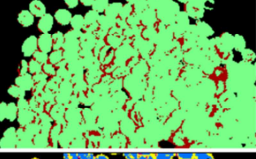
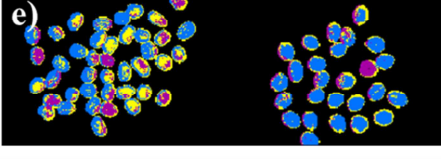
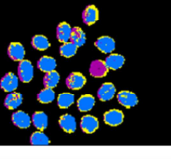
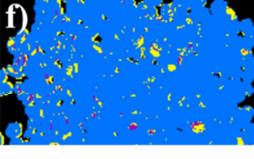
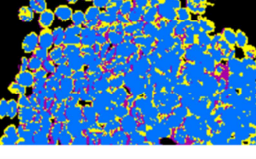
20 In fact, the predicted image obtained when AS training set was used to classify test image 1 (Figure  
 21 2.a), in which the pixels predicted as Arabica coffee are represented in red colour and the pixels

1 predicted as Robusta coffee are represented in green colour, shows that the classification  
2 performances are strongly influenced by the round shape of the beans. Actually, when dealing with  
3 the classification of objects using hyperspectral or multispectral imaging systems, the problem of  
4 the irregular shape of the samples is often present, since in practical applications the sample surface  
5 is not generally flat. This might be ascribed to a lower signal in the spectra or to the fact that the  
6 signal is mixed with the background signal.

7 The same model applied to test image 2 gives a higher *EFF* value, but the problem of the shape still  
8 occurs (Figure 2.b). In the case of the beans belonging to Robusta coffee (group of beans on the  
9 right) the misclassified pixels can be found mainly in correspondence of the edges of the beans.  
10 Furthermore, considering the beans belonging to Arabica coffee, some of those beans are  
11 misclassified (Figure 2.b).

12 Conversely, the shape effect is less evident when the same images are predicted with RS training set  
13 (Figures 2.c and 2.d). In particular, the results for test image 1 (Figure 2.c) show that one Robusta  
14 bean is clearly misclassified, and the same for some Arabica beans. Concerning test image 2, the  
15 major part of pixels is assigned to the correct class, even if some misclassifications occur at the  
16 edges of the beans (Figure 2.d). The differences in prediction on the test images of the two  
17 approaches are highlighted in Figure 2.e and Figure 2.f, which represent a difference image between  
18 the predictions made from AS training set and RS training set on the same image. In particular, the  
19 pixels correctly predicted with both approaches are represented in blue colour, the pixels  
20 misclassified with both approaches are represented in purple colour and those differently predicted  
21 (i.e., where only one of the two models have failed) are represented in yellow colour.

22

	TEST IMAGE 1		TEST IMAGE 2	
	Arabica	Robusta	Arabica	Robusta
<b>Predictions of model from AS training set</b> (red = predicted as Arabica; green = predicted as Robusta)	<b>a)</b> 		<b>b)</b> 	
<b>Predictions of model from RS training set</b> (red = predicted as Arabica; green = predicted as Robusta)	<b>c)</b> 		<b>d)</b> 	
<b>Difference between AS and RS predictions</b> (blue = both correct; purple = both misclassified; yellow = differently predicted)	<b>e)</b> 		<b>f)</b> 	

**Figure 2.** Filter simulation 1: predicted images of PLS-DA models built using AS training set (a, b) and RS training set (c, d), and difference images for test image 1 (e) and test image 2 (f).

Moreover, Table S-1 (Supplementary material) reports the percentage of correctly predicted pixels of Arabica and Robusta classes for both the test images. While the percentage of correctly predicted pixels for Arabica Coffee does not depend significantly upon the considered training set, the percentage of correctly predicted pixels for Robusta coffee is much higher when RS training set is used.

### 3.2 Filter Simulation 2

In Filter Simulation 2, the PLS-DA models were calculated using all the 32 descriptors listed in Table 1; the results obtained with the AS and the RS training sets are reported in the third and fourth column of Table 2, respectively. Interestingly, for both the AS and RS training sets only three LVs have been selected.

As observed in Filter Simulation 1, the classification model built using AS training set shows better performances in calibration and cross validation than the one built on RS training set. However, the model based on RS training set shows better results for the prediction both of test set and of test image 2. In particular, an *EFF* value equal to 100% is obtained when the classification model built

1 using RS training set is used to predict the test set. The performances in the prediction of test image  
2 1 are essentially the same in terms of efficiency values, but AS training set gives a higher  
3 percentage of Arabica pixels which are correctly predicted (76.8%), while RS training set gives a  
4 higher percentage of correctly predicted pixels for Robusta class (86.0%).

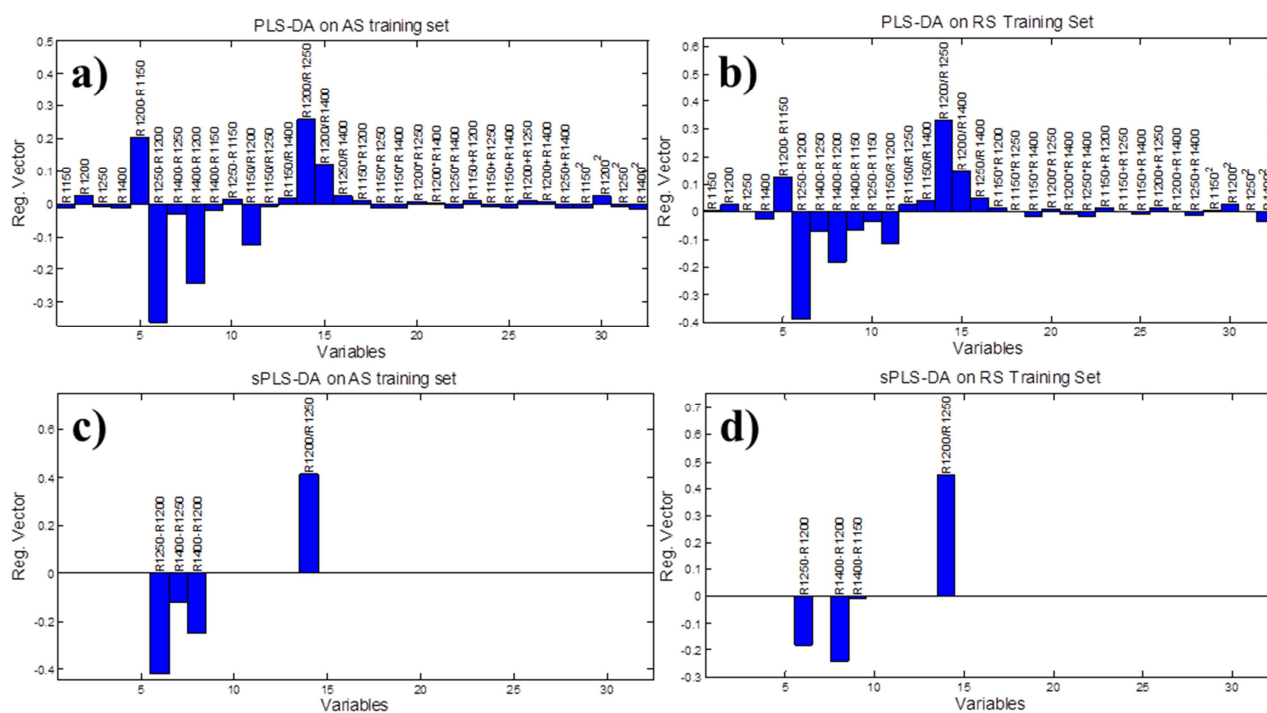
5 Comparing the results obtained for Filter Simulation 2 with those of Filter Simulation 1, it can be  
6 observed that, for the models calculated using AS training set, the use of additional descriptors led  
7 to a general increase of the *EFF* values; conversely, for the models calculated using RS training set,  
8 Filter Simulation 2 led to slightly worse results, except for the prediction of the test set.

9 Concerning the pixel-level predictions, Figures S-4.a and S-4.b (Supplementary material) report the  
10 predictions obtained using AS training set for test image 1 and test image 2, respectively. In both  
11 the images the problem of the round shape of the beans is less evident than in the corresponding  
12 predicted images obtained from Filter Simulation 1, even if many misclassifications still occur in  
13 correspondence to the edges and to the centre cut of the beans. Moreover, also the percentage of  
14 pixels correctly predicted generally increases from Filter Simulation 1 to Filter Simulation 2 for  
15 both test images mainly considering class Arabica.

16 Concerning the model built using RS training set, Figures S-4.c and S-4.d (Supplementary material)  
17 report the predictions of test image 1 and test image 2, which look very similar to those obtained in  
18 Filter Simulation 1.

19 The corresponding regression vectors have been reported in Figures 3.a and 3.b, respectively. For  
20 both the classification models, single reflectance values, sums and products have low influence,  
21 since the absolute values of the corresponding regression coefficients are small. The largest absolute  
22 values of the regression coefficients are instead observed for some differences and ratios; in  
23 particular, R1250-1200 and R1200/R1250 are the descriptors showing the highest contribution for  
24 both the models. More in general, the two regression vectors are quite similar to each other,  
25 suggesting that - independently of the considered training set - the information useful to  
26 discriminate between Arabica and Robusta coffee is found within the same subset of descriptors.

1



2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

**Figure 3.** PLS-DA regression vectors of the classification models built on AS training set (a) and RS training set (b) considered in filter simulation 2; and regression vectors of sPLS-DA models calculated on AS training set (a) and RS training set (b) in filter simulation 3.

### 3.3 Filter Simulation 3

In order to better identify the most relevant descriptors involved in the discrimination between Arabica and Robusta coffee beans and to reduce the number of variables used in the classification models, feature selection by sPLS-DA was then applied both to AS and to RS training sets.

Figure S-5 (Supplementary material) reports the response surfaces for AS and for RS training sets (Figures S-5.a and S-5.b, respectively) of the *EFF* values estimated in cross-validation as a function of the number of LVs and of the number variables selected for each LV. In both the cases, the optimal conditions were reached with 2 LVs and 2 variables for each component, which correspond to regression vectors including 4 selected variables, as reported in Figure 3.c for AS training set and in Figure 3.d for RS training set. Both sparse models converge on the selection of the three descriptors showing the highest absolute values of the corresponding regression coefficients, i.e.,

1 R1250-R1200, R1400-R1200 and R1200/R1250; moreover, R1400-R1250 and R1400-R1150 are  
2 additionally selected in the sparse model built on AS training set and RS training set, respectively.

3 The results of the sparse models are reported in the last two columns of Table 2. Similarly to the  
4 results of Filter Simulations 1 and 2, also for Filter Simulation 3 the *EFF* values obtained in  
5 calibration and in cross-validation for RS training set are lower than those obtained for AS training  
6 set, while higher performances have been obtained in prediction, both at the image-level and at the  
7 pixel-level. In particular, the use of the RS training set led to an *EFF* value equal to 100% for the  
8 prediction of the test set at the image-level; the corresponding sLV1 and sLV2 score plot is reported  
9 in Figure S-6 (Supplementary material) which shows a distinct separation of the samples belonging  
10 to Arabica and Robusta coffee species.

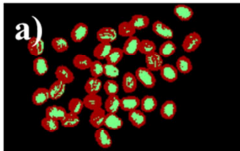
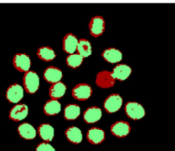
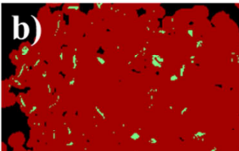
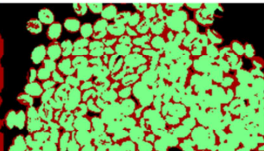
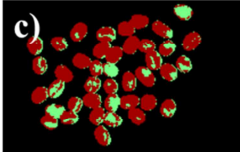
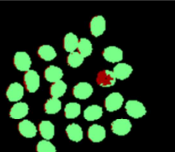
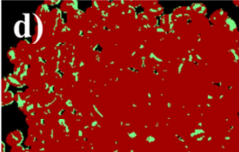
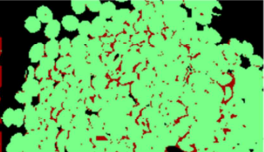
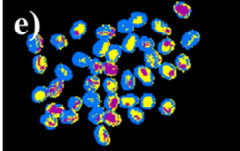
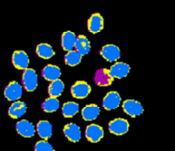
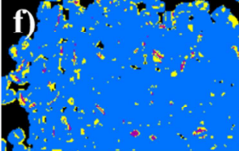
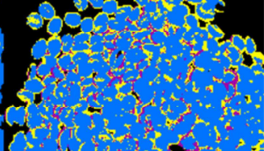
11 In general, compared to the use of the 4 reflectance values of the filters (Filter Simulation 1), the  
12 selection by sPLS-DA of the most relevant descriptors allowed to obtain more parsimonious models  
13 (2 LVs instead of 4 LVs), and to gain a significant increase of the classification performances in  
14 prediction, in particular for the model calculated using the RS training set.

15 As far as the prediction at the pixel-level is concerned, the comparison of the predicted images  
16 obtained from AS training set (Figures 4.a and 4.b for test image 1 and test image 2, respectively)  
17 and from RS training set (Figures 4.c and 4.d for test image 1 and test image 2, respectively) shows  
18 that the sparse model built using AS training set is much more sensitive to the round shape of the  
19 coffee beans, this fact being particularly evident in the case of test image 1. Moreover, in the  
20 prediction of test image 1 for both models more misclassifications occur for Arabica coffee beans  
21 (on the left) than for Robusta coffee beans (on the right). Indeed, for both models the percentage of  
22 correctly classified Robusta coffee pixels is greater than the percentage of correctly classified  
23 Arabica coffee pixels (Table S-1 in Supplementary material).

24 As far as test image 2 is concerned, even if the overall classification of the beans is correct for both  
25 models, the sparse classification model built using RS training set is less sensitive to edge effects  
26 and to the round shape of the beans (Figure 4.d). This is also evident from the difference image

1 between the predictions made with AS training set and RS training set, reported in Figures 4.e and  
 2 4.f, respectively: considering test image 1, one Robusta bean and four Arabica beans are always  
 3 misclassified (purple colour), while the pixels differently predicted (yellow colour) are mainly  
 4 ascribable to the fact that the model built using AS training set is more influenced by the shape of  
 5 the beans. Conversely, in test image 2 the overall classification of the beans is correct for both  
 6 classes; also in this case the differences are mainly due to a higher sensitivity to shape-related  
 7 effects in the model built using AS training set, as it is also confirmed by the values reported in  
 8 Table S-1 (Supplementary material).

9

	TEST IMAGE 1		TEST IMAGE 2	
	Arabica	Robusta	Arabica	Robusta
<b>Predictions of model from AS training set</b> (red = predicted as Arabica; green = predicted as Robusta)				
<b>Predictions of model from RS training set</b> (red = predicted as Arabica; green = predicted as Robusta)				
<b>Difference between AS and RS predictions</b> (blue = both correct; purple = both misclassified; yellow = differently predicted)				

10

11

12

13

**Figure 4.** Filter simulation 3: predicted images of sPLS-DA models built using AS training set (a, b) and RS training set (c, d) and difference images for test image 1 (e) and test image 2 (f).

14

15

16

17

18

19

In order to better investigate the influence of shape-related effects, the Hotelling's  $T^2$  values and the Q residuals of the sPLS-DA models discussed in this Section have also been considered. Actually, the pixels falling outside the 95% confidence limits are mainly placed at the edges of the beans, as reported in Figure S-7 of Supplementary material for the sPLS-DA model calculated with RS training set. In this case, only a small percentage of pixels was detected as outlier, thus the elimination of those pixels would have not significantly increased the classification performances.

1 However, it has to be highlighted that the implementation of outlier statistics in the classification  
2 model could be considered for a further optimization, giving the possibility to detect also potential  
3 foreign objects, such as metals, stones, sticks or soil residuals.

4

5

#### 6 **4 CONCLUSIONS**

7 This paper is aimed at investigating the issues concerning the implementation of a multispectral  
8 imaging system, starting from the outcome of spectral feature selection performed using a  
9 hyperspectral system, in order to discriminate between Arabica and Robusta coffee species. In  
10 particular, the simulations reported in the present work were done not only considering the single  
11 reflectance values of the filters, but also exploring systematically linear and non-linear relationships  
12 between bands, and selecting the relevant descriptors involved in the classification, which led to a  
13 significant increase of the classification performances in prediction.

14 Moreover, attention was also paid to the proper construction of a representative training set by  
15 comparing two different approaches, which consisted in the use of average image spectra and of  
16 randomly selected spectra; the latter one led to better classification results, mostly in the case of  
17 predictions at the pixel-level, since it allowed to consider also the spatial variability within each  
18 image.

19 In the specific case, the use of a representative training set of randomly selected spectra allowed to  
20 better account for the round shape of the green coffee beans, while the selection of the most  
21 effective combinations of channels allowed to explore the descriptors which better reflect the  
22 chemical differences between Arabica and Robusta green coffee.

23 In particular, the proper combination of four NIR band-pass filters and the use of multivariate  
24 statistical analysis allowed to achieve classification performances comparable to those obtained  
25 with the hyperspectral data, i.e. considering the full NIR spectrum. Moreover, it has to be  
26 highlighted that the calculation of the 32 descriptors from the outputs of the band-pass filters and

1 the choice of the relevant ones involved in discrimination are conducted off-line during the  
2 optimization of the classification model. For the real-time implementation of the model only the  
3 four selected descriptors have to be calculated, which is much faster than acquiring and elaborating  
4 the whole spectrum (150 spectral variables) of the original hyperspectral imaging system. Based on  
5 these results, a filter-based multispectral imaging system can be easily implemented in order to  
6 obtain a fast and inexpensive tool suitable for on-line monitoring of green coffee, allowing to  
7 prevent counterfeits.

8 In general, the proposed approach demonstrated the possibility of having a clear idea of the  
9 classification performances that can be reached using a multispectral system, much time before the  
10 system itself is actually constructed.

## 11

## 12 **ACKNOWLEDGMENTS**

13 The authors wish to thank Luigi Bellucci (Caffè Molinari S.p.A) for providing the coffee samples  
14 and technical support.

15

1 **REFERENCES**

- 2 [1] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O.L. García-Navarrete, J. Blasco, Recent  
3 advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food*  
4 *Bioprocess Tech.* 5:4 (2012) 1121-1142.
- 5 [2] G. Elmasry, M. Kamruzzaman, D.W Sun, P. Allen, Principles and applications of hyperspectral  
6 imaging in quality evaluation of agro-food products: a review, *CRC Cr. Rev. Food Sci.* 52:11  
7 (2012) 999-1023.
- 8 [3] A.A. Gowen, C. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging—an  
9 emerging process analytical tool for food quality and safety control. *Trends Food Sci. Tech.* 18:12  
10 (2007) 590-598.
- 11 [4] J.M. Amigo, I. Martí, A. Gowen, A. Hyperspectral imaging and chemometrics: a perfect  
12 combination for the analysis of food structure, composition and quality. In: F. Marini (Ed), *Data*  
13 *Handling in Science and Technology.* Vol 28, 2013, pp. 343-370.
- 14 [5] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: reducing data  
15 size without loss of useful information, *Anal. Chim. Acta* 802 (2013) 29-39.
- 16 [6] C. Ferrari, G. Foca, R. Calvini, A. Ulrici, Fast exploration and classification of large  
17 hyperspectral image datasets for early bruise detection on apples, *Chemometr. Intell. Lab.* 146  
18 (2015) 108-119.
- 19 [7] R. Calvini, A. Ulrici, J.M. Amigo, Practical comparison of sparse methods for classification of  
20 Arabica and Robusta coffee species using near infrared hyperspectral imaging, *Chemometr. Intell.*  
21 *Lab.* 146 (2015) 503-511.
- 22 [8] D. Lorente, N. Aleixos, J. Gomez-Sanchis, S. Cubero, O.L. Garcia-Navarrete, J. Blasco, Recent  
23 advances and applications of hyperspectral imaging for fruit and vegetable quality assessment, *Food*  
24 *and Bioprocess Technology*, 5:4 (2012) 1121-1142.
- 25 [9] P.M. Mehl, K. Chao, M. Kim, Y.R. Chen, Detection of defects on selected apple cultivars using  
26 hyper spectral and multispectral image analysis, *Appl. Eng. Agric.* 18:2 (2002) 219.
- 27 [10] W. Huang, J. Li, Q. Wang, L. Chen, Development of a multispectral imaging system for online  
28 detection of bruises on apples, *J. Food Eng.* 146 (2015) 62-71.
- 29 [11] O. Kleynen, V. Leemans, M.F. Destain, Development of a multi-spectral vision system for the  
30 detection of defects on apples, *J. Food Eng.* 69:1 (2005) 41-49.
- 31 [12] J. Qin, T.F. Burks, X. Zhao, N. Niphadkar, M.A. Ritenour, Development of a two-band  
32 spectral imaging system for real-time citrus canker detection. *J. Food Eng.* 108:1 (2012) 87-93.
- 33 [13] M.O. Ngadi, L. Liu, Hyperspectral image processing techniques, In: D.W. Sun (Ed.),  
34 *Hyperspectral imaging for food quality analysis and control*, 2010, pp 99-127.
- 35 [14] B. Park, K.C. Lawrence, W.R. Windham, D.P. Smith, Performance of hyperspectral imaging  
36 system for poultry surface fecal contaminant detection, *J. Food Eng.* 75:3 (2006) 340-348.
- 37 [15] M.J. Martín, F. Pablos, A.G. González, Discrimination between arabica and robusta green  
38 coffee varieties according to their chemical composition, *Talanta* 46:6 (1998) 1259-1264.
- 39 [16] F. Carrera, M. León-Camacho, F. Pablos, A.G. González, Authentication of green coffee  
40 varieties according to their sterolic profile, *Anal. Chim. Acta* 370:2 (1998) 131-139.
- 41 [17] C. Sanz, L. Maeztu, M.J. Zapelena, J. Bello, C. Cid, Profiles of volatile compounds and  
42 sensory analysis of three blends of coffee: influence of different proportions of Arabica and Robusta  
43 and influence of roasting coffee with sugar, *J. Sci. Food Agr.* 82:8 (2002) 840-847.

- 1 [18] R.C. Alves, S. Casal, M.R. Alves, M.B. Oliveira, Discrimination between arabica and robusta  
2 coffee species on the basis of their tocopherol profiles, *Food Chem.* 114 (2009) 295-299.
- 3 [19] R.M. El-Abassy, P. Donfack, A. Materny, Discrimination between Arabica and Robusta green  
4 coffee using visible micro Raman spectroscopy and chemometric analysis, *Food Chem.* 126:3  
5 (2011) 1443-1448.
- 6 [20] R. Garrett, B.G. Vaz, A.M.C. Hovell, M.N. Eberlin, C.M. Rezende, Arabica and Robusta  
7 coffees identification of major polar compounds and quantification of blends by direct-infusion  
8 electrospray ionization-mass spectrometry. *Journal of Agricultural and Food Chemistry*, 60:17  
9 (2012) 4253-4258.
- 10 [21] C. Martellosi, E.J. Taylor, D. Lee, G. Graziosi, P. Donini, DNA extraction and analysis from  
11 processed coffee beans, *Journal of Agricultural and Food Chemistry*, 53:22 (2005) 8432-8436.
- 12 [22] A.T. Toci, A. Farah, H.R. Pezza, L. Pezza, Coffee adulteration: More than two decades of  
13 research, *Critical Reviews in Analytical Chemistry*, 46:2 (2016) 83-92.
- 14 [23] Y.B. Monakhova, W. Ruge, T. Kuballa, M. Ilse, O. Winkelmann, B. Diehl, D.W. Lachenmeier,  
15 Rapid approach to identify the presence of Arabica and Robusta species in coffee using  $^1\text{H}$  NMR  
16 spectroscopy, *Food Chem.* 182 (2015) 178-184.
- 17 [24] F. Wei, K. Furihata, M. Koda, F. Hu, R. Kato, T. Miyakawa, M. Tanokura,  $^{13}\text{C}$  NMR-based  
18 metabolomics for the classification of green coffee beans according to variety and origin, *J. Agr.*  
19 *Food Chem.* 60:40 (2012) 10118-10125.
- 20 [25] M. T. Gil-Agusti, N. Campostrini, L. Zolla, C. Ciambella, C. Invernizzi, P.G. Righetti, Two-  
21 dimensional mapping as a tool for classification of green coffee bean species, *Proteomics* 5:3  
22 (2005) 710-718.
- 23 [26] S. Buratti, N. Sinelli, E. Bertone, A. Venturello, E. Casiraghi, F. Geobaldo, Discrimination  
24 between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy,  
25 electronic nose and electronic tongue analysis, *J. Sci. Food Agr.* 95:11 (2015) 2192-2200.
- 26 [27] E.K. Kemsley, S. Ruault, R.H. Wilson, Discrimination between *Coffea arabica* and *Coffea*  
27 *canephora* variant robusta beans using infrared spectroscopy. *Food Chem.* 54:3 (1995) 321-326.
- 28 [28] J.R. Santos, M.C. Sarraguça, A.O. Rangel, J.A. Lopes, Evaluation of green coffee beans  
29 quality using near infrared spectroscopy: A quantitative approach. *Food Chem.* 135:3 (2012) 1828-  
30 1835.
- 31 [29] A.J. Myles, T.A. Zimmerman, S.D. Brown, Transfer of multivariate classification models  
32 between laboratory and process near-infrared spectrometers for the discrimination of green Arabica  
33 and Robusta coffee beans. *App. Spectrosc.* 60:10 (2006) 1198-1203.
- 34 [30] I. Esteban-Diez, J.M. González-Sáiz, C. Pizarro, An evaluation of orthogonal signal correction  
35 methods for the characterisation of arabica and robusta coffee varieties by NIRS, *Anal. Chim. Acta*  
36 514:1 (2004) 57-67.
- 37 [31] I. Esteban-Diez, J.M. González-Sáiz, C. Sáenz-González, C. Pizarro, Coffee varietal  
38 differentiation based on near infrared spectroscopy, *Talanta*, 71:1 (2007) 221-229.
- 39 [32] A.G. Fiore, R. Romaniello, G. Peri, C. Severini, Quality assessment of roasted coffee blends  
40 by hyperspectral image analysis. In *Proceedings of 22nd International Conference on Coffee*  
41 *Science*, Campinas, Brazil, 2008.
- 42 [33] A. Backhaus, F. Bollenbeck, U. Seiffert, High-throughput quality control of coffee varieties  
43 and blends by artificial neural networks and hyperspectral imaging. In *Proceedings of the 1st*  
44 *International Congress on Cocoa, Coffee and Tea*, CoCoTea, 2011.

- 1 [34] S. Chevallier, D. Bertrand, A. Kohler, P. Courcoux, Application of PLS-DA in multivariate  
2 image analysis, *J. Chemometr.* 20:5 (2006) 221.
- 3 [35] M.A.F. de la Ossa, C. García-Ruiz, J.M. Amigo, Near infrared spectral imaging for the analysis  
4 of dynamite residues on human handprints, *Talanta*, 130 (2014) 315-321.
- 5 [36] K.A. Lê Cao, S. Boitard, P. Besse, Sparse PLS discriminant analysis: biologically relevant  
6 feature selection and graphical displays for multiclass problems, *BMC bioinformatics*, 12(1) (2011)  
7 253.
- 8 [37] J. Burger, P. Geladi, Hyperspectral NIR image regression part II: dataset preprocessing  
9 diagnostics, *J. Chemometr.* 20:3 (2006) 106-119.
- 10 [38] J. Burger, P. Geladi, Hyperspectral NIR image regression part I: calibration and correction, *J.*  
11 *chemometr.* 19:5 (2005) 355-363.
- 12 [39] A. Ulrici, S. Serranti, C. Ferrari, D. Cesare, G. Foca, G. Bonifazi, Efficient chemometric  
13 strategies for PET-PLA discrimination in recycling plants using hyperspectral imaging,  
14 *Chemometr. Intell. Lab.*, 122 (2013) 31-39.
- 15 [40] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and  
16 classification with application to chemometrics, *J. Chemometr.* 26 (2012) 42-51.
- 17 [41] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modelling, *Chemometr.*  
18 *Intell. Lab.* 119 (2012) 21-31.
- 19 [42] T.M. Cover, P.E. Hart, Nearest neighbour pattern classification, *IEEE T. Inform. Theory*, 13:1  
20 (1967) 21-27.
- 21 [43] J.S. Shenk, J.J. Workman, M.O. Westerhaus, Application of NIR spectroscopy to agricultural  
22 products. In D.A. Burns, E.W. Ciurczak (Eds.), *Handbook of Near Infrared Analysis*, Third Edition,  
23 Boca Raton: CRC Press, 2008, pp. 347-386.
- 24 [44] NIR Bandpass & Laser Line Filters: 700 - 1650 nm Center Wavelength. URL  
25 [http://www.thorlabs.de/newgrouppage9.cfm?objectgroup\\_id=1000](http://www.thorlabs.de/newgrouppage9.cfm?objectgroup_id=1000) . Accessed 19.08.15.
- 26 [45] M. Calderisi, A. Ulrici, S. Sinisalo, J. Uotila, R. Seeber, Simulation of an experimental  
27 database of infrared spectra of complex gaseous mixtures for detecting specific substances. The  
28 case of drug precursors. *Sensor. Actuat. B: Chem.* 193 (2014) 806-814.
- 29 [46] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant  
30 partial least-squares binary classification, *Chemometr. Intell. Lab.* 95:2 (2009) 122-128.
- 31 [47] E. Szymanska, E. Brodrick, M. Williams, A.N. Davies, H.J. Van Manen, L.C.M Buydens, Data  
32 size reduction strategy for the classification of breath and air samples using multicapillary column-  
33 ion mobility spectrometry, *Analytical Chemistry* 87:2 (2015) 869-875.

34

## 1 CAPTURES TO FIGURES

2 **Figure 1.** Average spectra calculated from each image of the training samples with selected regions  
3 highlighted (a), Gaussian profiles of the 4 filters considered (b) and resulting reflectance  
4 values obtained from the average training spectra (c).

5 **Figure 2.** Filter simulation 1: predicted images of PLS-DA models built using AS training set (a, b)  
6 and RS training set (c, d), and difference images for test image 1 (e) and test image 2 (f).

7 **Figure 3.** PLS-DA regression vectors of the classification models built on AS training set (a) and  
8 RS training set (b) considered in filter simulation 2; and regression vectors of sPLS-DA  
9 models calculated on AS training set (a) and RS training set (b) in filter simulation 3.

10 **Figure 4.** Filter simulation 3: predicted images of sPLS-DA models built using AS training set (a,  
11 b) and RS training set (c, d) and difference images for test image 1 (e) and test image 2  
12 (f).

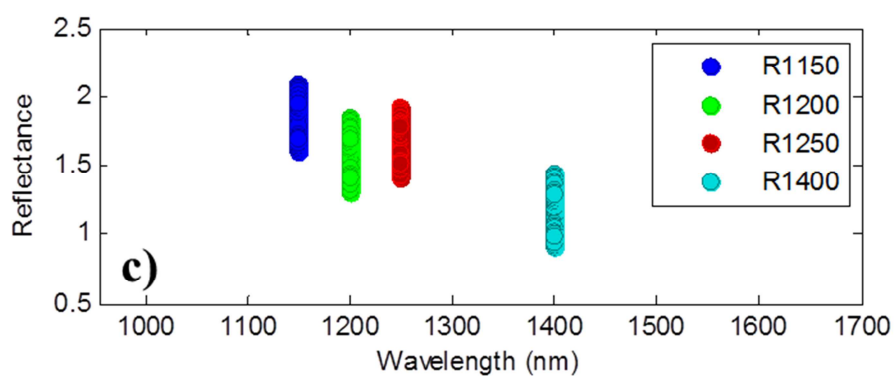
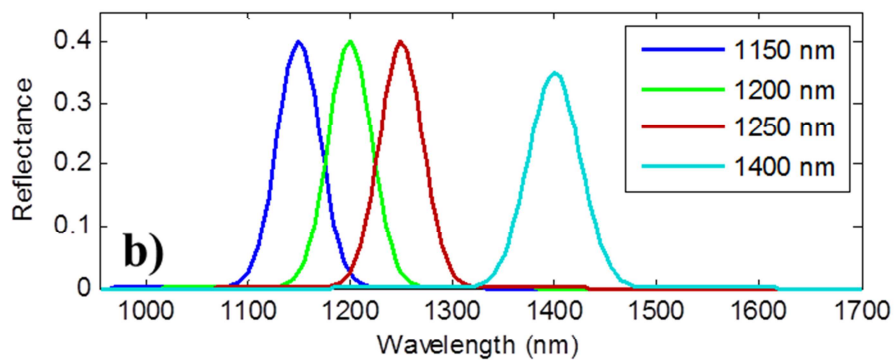
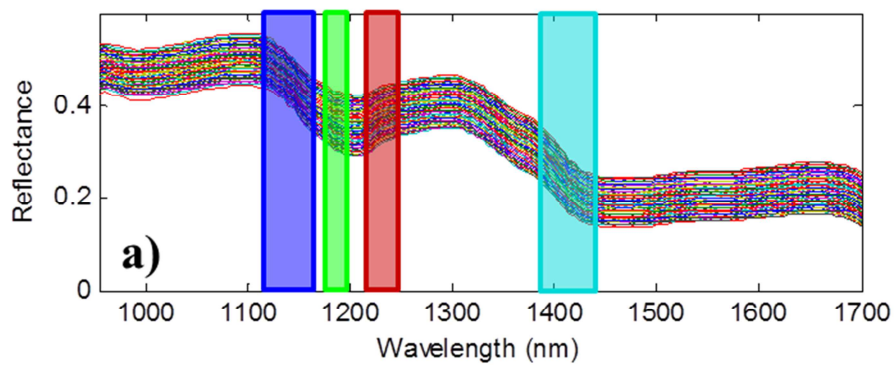
13

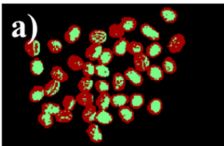
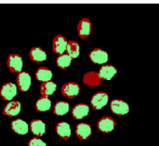
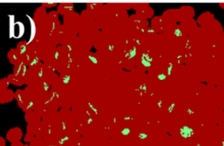
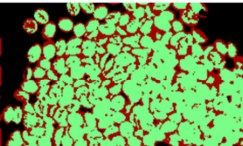
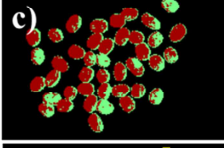
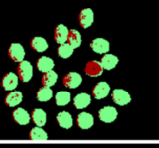
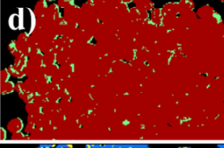
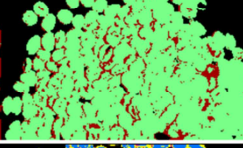
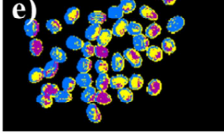
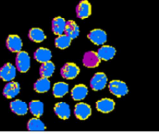
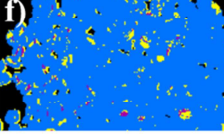
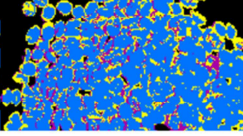
**Table 1.** List of the 32 descriptors considered.

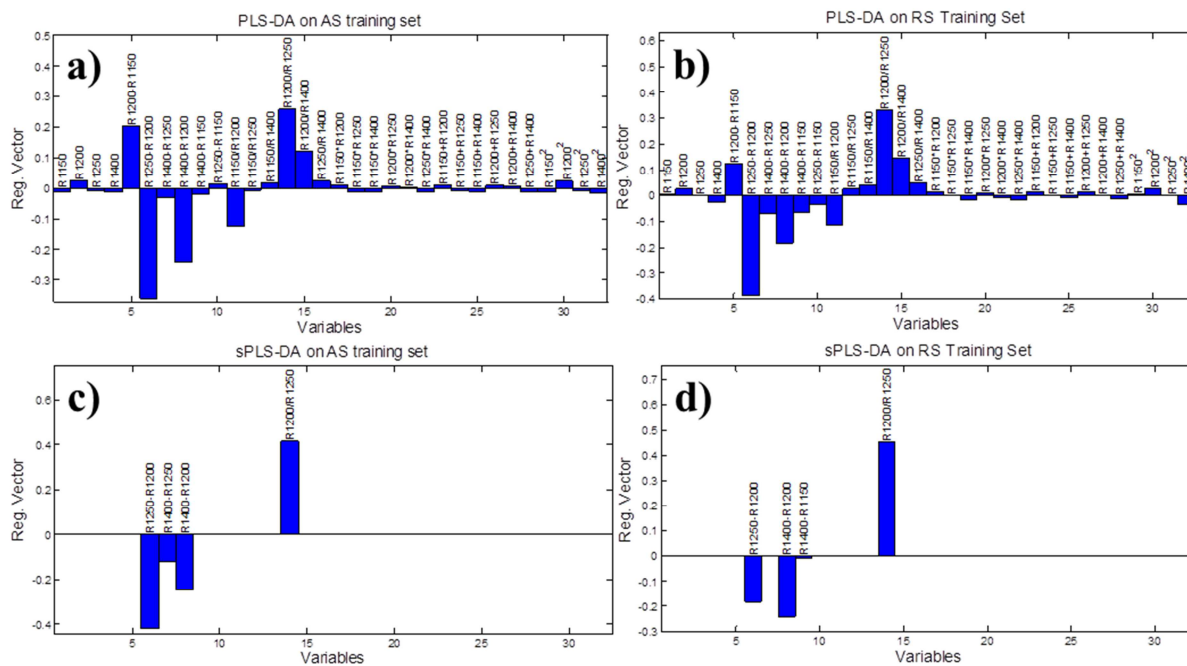
Single Filters	Differences	Ratios	Products	Sums
R1150	R1200-R1150	R1150/R1200	R1150*R1200	R1150+R1200
R1200	R1250-R1200	R1150/R1250	R1150*R1250	R1150+R1250
R1250	R1400-R1250	R1150/R1400	R1150*R1400	R1150+R1400
R1400	R1400-R1200	R1200/R1250	R1200*R1250	R1200+R1250
R1150 <sup>2</sup>	R1400-R1150	R1200/R1400	R1200*R1400	R1200+R1400
R1200 <sup>2</sup>	R1250-R1150	R1250/R1400	R1250*R1400	R1250+R1400
R1250 <sup>2</sup>				
R1400 <sup>2</sup>				

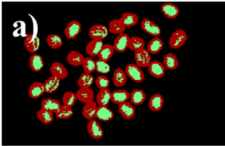
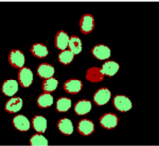
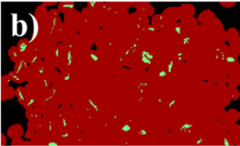
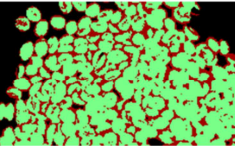
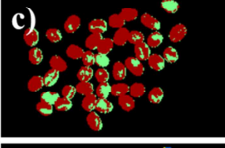
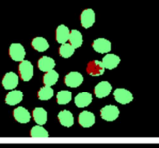
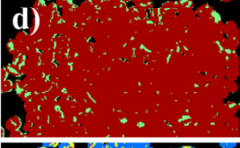
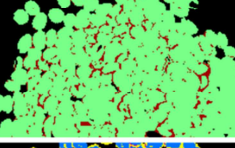
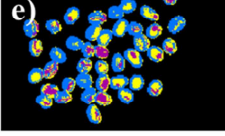
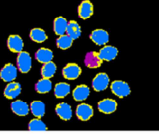
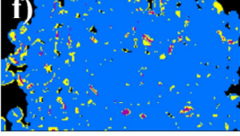
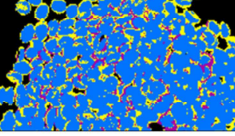
**Table 2.** Results obtained using AS training set and RS training set in filters simulation1, filter simulation 2 and filter simulation 3; the classification efficiency values are referred to the results in calibration ( $EFF_{CAL}$ ), cross-validation ( $EFF_{CV}$ ), prediction of the test set at the image-level ( $EFF_{TEST}$ ), and of test image 1 and test image 2 at the pixel-level ( $EFF_{IMG1}$  and  $EFF_{IMG2}$ , respectively).

	Filter Simulation 1		Filter Simulation 2		Filter Simulation 3	
	AS tr. set	RS tr. set	AS tr. set	RS tr. set	AS tr. set	RS tr. set
<b>N° variables</b>	4	4	32	32	4	4
<b>LVs</b>	4	4	3	3	2	2
<b><math>EFF_{CAL}</math></b>	99.1	81.3	100.0	80.3	99.6	78.4
<b><math>EFF_{CV}</math></b>	97.3	79.9	99.3	78.7	97.3	78.0
<b><math>EFF_{TEST}</math></b>	94.0	94.9	97.5	100.0	98.3	100.0
<b><math>EFF_{IMG1}</math></b>	65.1	74.0	71.1	71.0	69.5	83.9
<b><math>EFF_{IMG2}</math></b>	83.9	92.2	84.3	92.1	85.6	93.1



	TEST IMAGE 1		TEST IMAGE 2	
	<i>Arabica</i>	<i>Robusta</i>	<i>Arabica</i>	<i>Robusta</i>
<b>Predictions of model from AS training set</b> (red = predicted as <i>Arabica</i> ; green = predicted as <i>Robusta</i> )	<b>a)</b> 		<b>b)</b> 	
<b>Predictions of model from RS training set</b> (red = predicted as <i>Arabica</i> ; green = predicted as <i>Robusta</i> )	<b>c)</b> 		<b>d)</b> 	
<b>Difference between AS and RS predictions</b> (blue = both correct; purple = both misclassified; yellow = differently predicted)	<b>e)</b> 		<b>f)</b> 	



	TEST IMAGE 1		TEST IMAGE 2	
	<i>Arabica</i>	<i>Robusta</i>	<i>Arabica</i>	<i>Robusta</i>
<b>Predictions of model from AS training set</b> (red = predicted as <i>Arabica</i> ; green = predicted as <i>Robusta</i> )	<b>a)</b> 		<b>b)</b> 	
<b>Predictions of model from RS training set</b> (red = predicted as <i>Arabica</i> ; green = predicted as <i>Robusta</i> )	<b>c)</b> 		<b>d)</b> 	
<b>Difference between AS and RS predictions</b> (blue = both correct; purple = both misclassified; yellow = differently predicted)	<b>e)</b> 		<b>f)</b> 	

ACCEPTED MANUSCRIPT

**Highlights**

- Fast discrimination between Arabica and Robusta is important for coffee industry
- Multispectral imaging can be very effective to this aim, but hard to implement
- Multispectral data were simulated from HSI data and NIR band-pass filters
- sPLS-DA was used for classification and selection of the relevant descriptors
- Classification performances were evaluated both at image-level and at pixel-level