INTRODUCTION

Arabica and Robusta are the two main coffee species used in the preparation of commercial beverages, both alone or in blends. Since Arabica coffee is more expensive and of better quality than Robusta, the availability of fast analytical methods for noninvasive monitoring of the raw materials is extremely important in order to correctly discriminate the two coffee species and to prevent adulterations.

Hyperspectral imaging (HSI) has showed a great potential for rapid and non-destructive inspection of several food matrices1. Despite many advantages of this technique, the implementation of HSI for real-time monitoring systems is currently limited by the high costs of hyperspectral cameras and by the data handling issues related to the huge amount of data contained in each hyperspectral image2. For these reasons, spectral feature selection is commonly performed on hyperspectral data in order to select few key wavelengths to be used for the implementation of cheaper and faster multispectral imaging systems3.

However, transferring the outcomes of variable selection performed on hyperspectral data to filter-based imaging systems is not straightforward. In fact, in the case of multispectral imaging each pixel only contains discrete diffuse reflectance values measured on each selected spectral interval and not a continuum spectrum. On the other hand, the limited number of multispectral channels limits the use of pre-processing and poses the difficulty of finding additional descriptors by calculating quantities derived from the outputs of the different channels and accounting for linear and non-linear relationships between spectral channels.

In this context, the present work aims at investigating the feasibility of implementing a classification model for filter-based multispectral data in order to discriminate Arabica and Robusta green coffee, starting from the results of a variables selection/classification model calculated on hyperspectral data. In particular, we focused on two different aspects: the selection of proper descriptors, i.e. combinations of channels, involved in the classification and the use of a representative training set.

KEYWORDS: Arabica and Robusta green coffee; Sparse methods; NIR hyperspectral imaging; Multispectral imaging; Variable selection.

Due to high costs and data handling issues, hyperspectral imaging is not easily implementable for real-time non-invasive applications. Spectral feature selection performed on hyperspectral data allows identifying key wavelengths for the development of cheaper and faster multispectral systems. The present work aims at investigating the issues related to the development of a multispectral imaging system, starting from variable selection performed with hyperspectral systems. Firstly, sparse Partial Least Squares-Discriminant Analysis (sPLS-DA) was applied to hyperspectral data in order to select the relevant spectral regions involved in the classification of Arabica and Robusta green coffee. Afterwards, four bandpass filters matching these regions were considered in order to simulate multispectral data used to calculate multivariate classification models. The proper definition of the training set and the selection of the most relevant combinations between spectral channels allowed reaching classification performances similar to those obtained with the hyperspectral data.

KEYWORDS: Arabica and Robusta green coffee; Sparse methods; NIR hyperspectral imaging; Multispectral imaging; Variable selection.
MATERIALS AND METHODS

In this study, 33 different green coffee batches (18 of Robusta and 15 of Arabica coffee) were considered, each batch corresponding to 60 kg of green coffee beans. From each batch, a representative amount of 500 g of beans was sampled and, from this amount, 12 hyperspectral images were acquired obtaining a dataset composed of 396 images (= 12 images × 33 batches).

The hyperspectral images were acquired using a desktop NIR Spectral Scanner (DV Optic) embedding a Specim N17E reflectance imaging spectrometer, coupled to a Xenics XEVA 2608 camera (320 × 256 pixels) and covering the 955-1700 nm spectral range with a spectral resolution of 5 nm.

Before further analysis, the hyperspectral images were preprocessed using an internal calibration in order to reduce the variability among images over time, and the pixel related to the background were removed from each image. More details about image preprocessing of this dataset can be found in Calvini et al. 2015.

The 33 coffee batches were randomly divided into 24 training batches (288 images), containing 11 Arabica and 13 Robusta coffee batches, and in 9 test batches (108 images), containing 4 Arabica and 5 Robusta coffee batches. The hyperspectral images of the samples belonging to the training batches were used to constitute two different training sets: Average Spectra (AS) training set, obtained by calculating the average spectra from each image of the training batches, and Random Spectra (RS) training set, obtained by randomly selecting 50 spectra from each image of the training batches. Since each hyperspectral image is composed of tens of thousands pixel spectra, AS and RS training sets denote two different strategies used to build a reduced but still representative training set from the huge amount of data contained in the images. In particular, AS training set is built under the assumption that the average spectrum is representative of the whole image (thus losing the information related to spatial variability), while RS training set allows reducing data size and also accounting for spatial variability within each image.

The hyperspectral images of the test batches were used to calculate a test set of average spectra from each image, for the evaluation of the classification performances at the image level, i.e. considering each image on the whole. In addition, two images of the test samples, one of Arabica coffee and one of Robusta coffee, were merged together in order to create a test image used to evaluate the classification models also at the pixel level, i.e. evaluating the class assignment of each single pixel.

Firstly, Partial Least Squared Discriminant Analysis (PLS-DA) and sparse PLS-DA (sPLS-DA) were applied to AS training set in order to perform classification and, for sPLS-DA, variable selection. The average spectra were preprocessed using Standard Normal Variate coupled with first derivative and mean center.

Subsequently, starting from the results of the variable selection/classification model calculated on hyperspectral data, 4 commercially available bandpass filters were considered (1150 nm, 1200 nm, 1250 nm, 1400 nm) as those showing the best match with the selected regions. The hyperspectral datasets were then converted into multispectral data in order to estimate the reflectance values that would be obtained using the considered filters. To this aim, the Gaussian-shaped profile of each filter was calculated from the filter properties (filter width at half maximum and percentage of transmission) provided by the commercial house. This profile was multiplied by the reflectance spectra measured with the hyperspectral system and the sum of the obtained values was calculated. In this manner, the hyperspectral data including 150 variables were transformed into multispectral data including 4 variables, where each variable is the reflectance value of a given multispectral channel (R1150, R1200, R1250, R1400). Furthermore, additional descriptors derived from the four reflectance values were also calculated, in order to consider linear and non-linear relationships between the spectral channels. In particular, the squared reflectance value of each channel and differences, ratios, products and sums between couples of channels were calculated. The resulting datasets were composed of 32 variables: four reflectance values of the filters and the additional descriptors. Classification was performed using both PLS-DA and sPLS-DA; the latter one was used in order to identify the most relevant descriptors. In this case, the classification models were built using both AS training set and RS training set, and the data were autoscaled before the analysis.

For both hyperspectral and multispectral data, cross validation was used in order to select the proper number of latent variables (LVs) and, for sPLS-DA, the number of variables to keep for each component. In particular, con-
tiguous block cross-validation was performed using 4 deletion groups, where each block contained the average spectra (for AS training set) or randomly selected spectra (for RS training set) of all the replicated and repeated images of 6 batches. The classification performances of the models were evaluated in terms of classification efficiency. Figure 1 shows the procedure followed in this study, from variable selection on hyperspectral data to the selection of proper conditions for the simulated multispectral system.

RESULTS AND DISCUSSION

Considering variable selection performed on hyperspectral data, different sPLS-DA models were calculated testing all the combinations between a number of latent variables ranging from 1 to 7 and a number of selected variables for each sLV ranging from 5 to 150. Based on the values of classification efficiency estimated in cross-validation, the optimal conditions were reached with 2 LVs and 10 variables for each component. This situation corresponds to a sparse regression vector with 20 non-zero variables out of 150.

Compared to PLS-DA, sPLS-DA allowed obtaining a more parsimonious model with similar classification performances (Table 1, Hyperspectral Data). In particular, for both PLS-DA and sPLS-DA an efficiency value equal to 100% was obtained in prediction on the test set, while sPLS-DA was slightly less performing at the pixel level prediction since the sparse model turned out to be more sensitive to the round shape of the beans.

The regions selected on the sparse loading vectors are related to the aromatic (1143 nm) and aliphatic (1195–1225 nm) C-H second overtone and to the O-H first overtone of aromatic (1410 nm) and aliphatic (1420) alcohols. The chemical relevance of these spectral regions in the classification of Arabica and Robusta green coffee was also confirmed by the results obtained with other sparse-based methods applied to the same dataset. The sparse loading vectors are reported in Figure 2 together with the spectra belonging to AS training set. Furthermore, in the same figure, the yellow bars indicate the spectral regions covered by the band-pass filters selected considering the results of sPLS-DA variable selection/classification model.

### Table 1. Classification results obtained from PLS-DA and sPLS-DA applied to hyperspectral and multispectral data

<table>
<thead>
<tr>
<th></th>
<th>Hyperspectral Data</th>
<th></th>
<th>Multispectral Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLS-DA</td>
<td>sPLS-DA</td>
<td>PLS-DA</td>
<td>sPLS-DA</td>
</tr>
<tr>
<td></td>
<td>AS tr. set</td>
<td>RS tr. set</td>
<td>AS tr. set</td>
<td>RS tr. set</td>
</tr>
<tr>
<td># var</td>
<td>150</td>
<td>20</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td># LVs</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>EFF&lt;sub&gt;cal&lt;/sub&gt;</td>
<td>100.0</td>
<td>99.4</td>
<td>100.0</td>
<td>80.3</td>
</tr>
<tr>
<td>EFF&lt;sub&gt;CV&lt;/sub&gt;</td>
<td>97.8</td>
<td>99.3</td>
<td>99.3</td>
<td>78.7</td>
</tr>
<tr>
<td>EFF&lt;sub&gt;test&lt;/sub&gt;</td>
<td>100.0</td>
<td>100.0</td>
<td>97.5</td>
<td>100.0</td>
</tr>
<tr>
<td>EFF&lt;sub&gt;MEG&lt;/sub&gt;</td>
<td>85.0</td>
<td>80.2</td>
<td>71.1</td>
<td>71.0</td>
</tr>
</tbody>
</table>

As far as the analysis of the multispectral data is concerned, different sPLS-DA models were tested on both AS and RS training sets considering all the combinations between a number of LVs from 1 up to 5 and a number of selected variables for each component ranging from 2 to 32. Also in this case classification efficiency in cross-
validation was considered in order to tune the model parameters; for both the training sets the proper conditions were chosen in correspondence to 2 LVs and 2 variables for each component, obtaining regression vectors with 4 selected variables. In particular, both sparse models converged to the selection of the three descriptors with the highest absolute values of the corresponding regression coefficients: R1250-R1200, R1400-R1200 and R1200-R1250. Moreover, R1400-R1250 and R1400-R1150 were selected in the sparse models built on AS and RS training sets, respectively.

Generally, compared to the PLS-DA model calculated considering all the 32 variables, the selection of the most relevant descriptors led to a significant increase of the classification performances in prediction, in particular for the model calculated using RS training set. Indeed, despite lower performances in calibration and cross-validation, RS training set led to higher efficiency values in prediction both for the test set and for the test image. In particular, an efficiency value equal to 100% was obtained when the sparse model built with RS training set was used for the prediction of the test set at the image level (Table 1, Multispectral Data).

Figure 3 shows the differences in the prediction of the test image between the sPLS-DA model calculated on hyperspectral data and the best sPLS-DA model calculated on multispectral data (i.e., the sPLS-DA model calculated from RS training set). The majority of the beans is correctly classified (white colour) in both cases even if one bean of Robusta coffee is always misclassified. Interestingly, the model calculated on hyperspectral data performed better for Arabica coffee, while the model calculated on multispectral data allowed to better classify Robusta coffee. In both cases the misclassifications are generally due to the round shape of the beans or to the presence of the centre cut.

**Figure 3.** Differences between predictions of the test image obtained with the best sPLS-DA models calculated using hyperspectral and multispectral data.

**CONCLUSIONS**

The proposed approach investigates the issues related to the development of a multispectral imaging system starting from the outcome of spectral feature selection achieved on hyperspectral systems. The performed simulations allowed to systematically evaluate also linear and non-linear relationships between spectral channels, in addition to the single reflectance values of the filters. In this manner, the combination of properly selected descriptors and the use of a representative training set allowed obtaining satisfactory results in the classification of Arabica and Robusta green coffee, with performances comparable to those obtained using the hyperspectral data both at the image level and at the pixel level. In general, the proposed approach allows assessing the actual potential of a multispectral imaging system before the system itself is built.

**Acknowledgments**

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

**References**