Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis

A. Giraudo[a], *S. Grassi*[b], *F. Savorani*[a*], *G. Gavoci*[a], *E. Casiraghi*[b], *F. Geobaldo*[a]

[a] Department of Applied Science and Technology, Polytechnic of Turin, Corso Duca degli Abruzzi 24, I-10129, Turin, Italy.

[b] Department of Food, Environmental and Nutritional Sciences, University of Milan, Via G. Celoria 2, I-20133, Milan, Italy.

*Corresponding author. E-mail: francesco.savorani@polito.it; Phone: 011 0904562

Authors' e-mail:

Alessandro Giraudo, alessandro.giraudo@polito.it

Silvia Grassi, silvia.grassi@unimi.it

Gentian Gavoci, gentian.gavoci@polito.it

Ernestina Casiraghi, ernestina.casiraghi@unimi.it

Francesco Geobaldo, francesco.geobaldo@polito.it

Abstract

In this work, near infrared (NIR) spectroscopy and multivariate data analysis were investigated as a fast and non-disruptive method to classify green coffee beans on continents and countries bases. FT-NIR spectra of 191 coffee samples, origin from 2 continents and 9 countries, were acquired by two different laboratories.

Laboratory-independent Partial Least Square-Discriminant Analysis and interval PLS-DA models were developed by following a hierarchical approach, i.e. considering at first the continent and then the country of origin as discrimination rule.

The best continent-based classification model was able to identify correctly more than 98% in prediction, whereas 100% of them were correctly predicted by the best country-based classification model. The inter-laboratory reliability of the proposed method was confirmed by McNemar test, since no significant differences ($P>0.05$) were found. Furthermore, a validation was performed predicting the spectral test set of a laboratory using the model developed by the other one.

Keywords

Geographical origin; Green coffee beans; NIR Spectroscopy; Chemometrics; Classification; Variable selection

1. Introduction

The sensory properties of coffee, one of the most popular beverages in the world, are deeply affected by the chemical composition of the raw coffee beans that is, in turn, highly related to their geographical growing regions (Esteban-Dìez et al., 2007). Coffee plantations are primarily found in the latitude between 20°N and 20°S, a geographical band encircling the world that is commonly known as "coffee belt". This band encompasses regions (belonging to Centre and South-America, Africa and Asia) roughly bounded by the tropics of Cancer and Capricorn, which offer optimal climate conditions for growing coffee plants. However, a great variance in coffee quality, taste and body is found within these world's regions (Anderson and Smith, 2002). Unavoidably, this variability aspect also implies considerable differences in terms of commercial values of the product (Teuber, 2010) and has led, in many cases, to common forms of frauds like mislabelling, i.e. disguising (not declaring) the right geographical or botanical origin of the coffee beans, and adulteration, i.e. selling cheaper coffee, or mix of coffees, as pure expensive species/categories (Alonso-Salces et al., 2009). As a consequence, coffee producers, as well as industrial manufacturers, are increasingly interested in protecting the market reputation from the aforementioned socioeconomic issues and have highly encouraged the development of efficient analytical methods able to evaluate the coffee authenticity (Huck et al., 2005).

In this context, during the last couple of decades, several analytical techniques, such as gas and liquid chromatography (Frega et al., 1995; Bicchi et al., 1997; Gonzalez et al., 2001; Casal et al., 2003; Martin et al., 2001; Carrera et al., 1998), mass spectrometry (Gil-Agusti et al., 2005) and nuclear magnetic resonance spectroscopy (Cagliani et al., 2013; Arana et al., 2015) were tested allowing gathering accurate information about the coffee composition, aiming at assessing its authenticity.

Unfortunately, these methods appear quite expensive, complex to use, often time consuming and requiring a sample preparation step before analysis that involves, in turn, the use of different kinds of chemical solvents. In that respect, alternative spectroscopic techniques, such as Raman and near-infrared (NIR) spectroscopy, provide a valid solution to overcome some of the abovementioned drawbacks, since they allow performing green, simple, fast and non-invasive analysis directly *in situ* (Marquetti et al., 2016). This is the reason why, over the years, they have found increasing application as routine analysis in different fields, such as biology, medicine, food science and technology (Ozaki et al., 2006).

For what it concerns the coffee production chain, Raman spectroscopy has been investigated for the chemical discrimination of Arabica and Robusta species (Rubayiza et al., 2005; El-Abassy et al.,

2011), while NIR spectroscopy has been successfully applied with different purposes, from the coffee varietal differentiation (Kemsley et al. 1995, Esteban-Dìez et al., 2007, Buratti et al., 2015) to the prediction of sensory properties of coffee beverage (Esteban-Díez et al., 2004; Ribeiro et al., 2011) and of the coffee roasting degree (Alessandrini et al., 2008; Bertone et al., 2016). Surprisingly, only a few research studies, regarding the use of NIR spectroscopy for the direct determination of the geographic origin of coffee, were found in literature. One of these studies focused on evaluating the variability of coffee samples cultivated in different regions of a single country (Marquetti et al., 2016). Another relevant study, performed by Medina et al., 2017, compared the ability of different spectroscopic techniques for the determination of the geographical origin of roasted coffee samples and was aimed at protecting the Colombian coffee origin. However, none of the proposed studies considered inter-laboratory results comparison, which is instead a relevant point when assessing an analytical method to be a reliable and accessible approach. Boix et al., 2012 evaluated the transfer of a NIR microscopy method for the detection of animal products in feedingstuffs performing Analysis of Variance (ANOVA) on the laboratories' duplicate analyses. A robust Ring test was performed involving 16 laboratories for the prediction of a significant number of wine quality parameters by Fourier transform-middle infrared spectrometry (Patz et al., 2004). However, to the best of our knowledge, this important aspect is usually underestimated in research studies dealing with food classification models based on NIR spectroscopy.

Furthermore, the NIR application in coffee field demands for the development of simplified devices (Pizarro et al., 2007; Calvini et al., 2017) reducing the costs in terms of money and time but leading to high quality results. The realization of *ad hoc* instruments passes through the identification of relevant spectral region, which can be achieved by variable selection strategies.

In the present work, NIR spectroscopy and multivariate data analysis were investigated as a fast and non-invasive method to classify green coffee beans cultivated in different world's continents and countries and to select relevant spectral features in vision of simplified instrument development.

In details, after data exploration by Principal Component Analysis (PCA), Partial Least Square-Discriminant Analysis (PLS-DA) algorithm was applied to develop a series of classification models. In this context, a hierarchical approach has been followed, considering at first the continent of origin and then the country of origin of the coffee samples as discrimination criteria. Moreover, interval Partial Least Square-Discriminant Analysis (iPLS-DA) algorithm was investigated to select the most informative spectral regions, and McNemar test was performed to compare the inter-laboratory model performance.

2. Materials and methods

2.1 Coffee samples

One hundred and ninety-one (191) samples of green coffee beans, representative of different geographical origins, were considered in the present study. Out of these, 88 coffee samples came from Centre-South America and 103 samples were grown in Asian countries; the details about the origin are provided in Table 1.

The coffee samples were representative of the main botanical species, i.e. *Robusta* and natural or washed *Arabica*. The producing countries and the species were chosen according to their relevance for the Italian coffee market.

The preliminary post-harvest processing operations on the green coffee beans, i.e. washing of cherry fruits and drying, were performed directly in the producing country.

Two replicates of sample, each one consisting of three hundred grams of coffee beans, were packed in vacuum in light barrier packaging material. The samples were prepared over a period of two months and delivered concurrently, within seven days from the preparation, in eight subsequent shipping to two different laboratories, i.e. the Department of Food, Environmental and Nutritional Sciences (UNIMI, University of Milan) and the Department of Applied Science and Technology (POLITO, Polytechnic of Turin).

Table 1

2.2 NIR spectra acquisition

NIR analysis was performed within 48 hours from the delivery of the coffee samples in both laboratories independently. The samples were analysed using two different Fourier Transform (FT)-NIR spectrometer (MPA, Bruker Optics, Ettlingen, Germany) provided with integrating sphere. The NIR spectra were collected at room temperature (~20° C) in diffuse reflectance mode directly on the entire coffee beans, hence without any kind of sample pre-treatment. Each sample was always carefully mixed before sampling eighty grams of coffee beans for each of the three replicates. During measurement, the sample holder was kept in rotation to collect more representative spectra on the whole sample. Spectral data were collected over the range 12500 to 3600 cm^{-1} (resolution, 8 cm^{-1}; scanner velocity, 10kHz; background, 256 scans; sample, 64 scans). Background scans were

performed using instrumental internal reference standard. OPUS software (v. 6.5, Bruker Optics, Ettlingen, Germany) was used for instrumental control and for spectra acquisition.

2.3 Multivariate data analysis

Chemometric elaboration of the NIR spectra was performed under Matlab environment (ver. 9.2, The Mathworks, Inc., Natick, USA) using the PLS toolbox (ver. 8.5, Eigenvector Research, Inc., USA) software package. The three replicates of NIR spectra collected on each coffee sample were averaged. Therefore, a global dataset, containing a unique, representative spectrum per sample, was obtained separately from both the laboratories and used independently for all the subsequent data elaboration steps.

Initially, different pre-processing methods were investigated aiming at eliminating potential artefacts from NIR spectra and correcting their nonlinear behaviour: Standard Normal Variate (SNV), Multiple Scattering Correction (MSC), as well as 1st or the 2nd derivative transformation using the Savitzky-Golay algorithm, applied alone or in combination with SNV or MSC. No matter the pre-treatment applied, all the data were subjected to mean centering prior to any multivariate data analysis.

2.3.1 Exploratory data analysis

Explorative data analysis was carried out by Principal Component Analysis (PCA), based on the singular value decomposition (SVD) algorithm (Wall et al., 2003), on all the datasets obtained by applying the different abovementioned pre-processing techniques. The optimal number of Principal Components (PCs) to be retained in the PCA model was chosen from the analysis of the corresponding scree plot (Bro and Smilde, 2014).

2.3.2 PLS-DA classification models

The classification models were developed using Partial Least Square – Discriminant Analysis (PLS-DA) algorithm (Wold et al., 1984).

The samples were divided into the proper number of classes according to the investigated discrimination criteria. In the first case, i.e. the continent-based discrimination, a vector of two *a-priori* classes was considered, where the first class referred to the American coffee samples and the other class referred to the Asian samples.

In the second case, i.e. the country-based discrimination, five classes were considered. Four of them corresponded to the four most representative producing countries with respect to the total number of samples reported in Table 1, i.e. Brazil (38), Honduras (27), India (65) and Vietnam (31). Class 5 groups 30 samples belonging to different countries, beings each country samples' number not enough to build an independent class maintaining a balanced design. In detail, this class includes samples produced in Guatemala (11), Colombia (8), Costa Rica (2), Nicaragua (2) and Indonesia (7), respectively.

Approximately the 75% of each sample dataset (144 samples, belonging to six deliveries) was used as calibration set and the remaining 25% (47 samples, belonging to two deliveries) was used as external validation set.

Models were tested internally by cross validating the calibration set using six cancellation groups, each group corresponding to one of the six deliveries. Each time, the samples included in the same cancellation group formed the prediction set, while all the other ones were used to build the model. Subsequently, the prediction capability of the achieved models was validated on the external test set.

The classification performance of the PLS-DA models was evaluated by comparing the actual (or reference) class to the class predicted by the model. In this context, four different conditions may occur based on the model prediction, since the samples can result "true positive" (TP), "true negative" (TN), "false positive" (FP) or "false negative" (FN) (Szymanska et al., 2012; Ballabio and Consonni, 2013). A series of statistical parameters were then calculated for all the class separately. They are:

- Sensitivity (*SENS*, Equation 1), which expresses the model capability to correctly recognize samples belonging to the considered class;
- Specificity (*SPEC*, Equation 2), which describes the model capability to correctly reject samples belonging to all the other classes;
- Efficiency (*EFF*), calculated as the geometric mean of *SPEC* and *SENS* (Equation 3).

$$SENS = \frac{TP}{(TP+FN)} \qquad \textbf{(Eq. 1)}$$

$$SPEC = \frac{TN}{(TN+FP)} \qquad \textbf{(Eq. 2)}$$

$$EFF = \sqrt{(SPEC * SENS)} \qquad \textbf{(Eq. 3)}$$

All these parameters can assume values between 0 (0%) and 1 (100%) and were calculated referring to the calibration (TRN, CAL), to the cross-validation of the training set (TRN, CV), and to the prediction of the external test set (TST, PRED).

Eventually, the classification results obtained were compared in terms of prediction performance by applying the 'testcholdout' function implemented in Matlab (v. 2016a, The Mathworks, Inc., Natick, USA), which performs one-tailed, mid P-value McNemar test (Fagerlan et al., 2013), a particular case of Fisher's sign test that verifies if two models have the same error rate (Roggo et al., 2003). For further details about the statistical test at hand, the reader is referred to Grassi et al. (2018a). In our case, the algorithm has been implemented to assess which data pre-treatment technique and/or combination of them gave the best results. Moreover, the test allowed to assess the inter-laboratory accuracy; that is whether the accuracy of the classification models obtained using the spectra collected with two different instruments by the same brand, but located in two different laboratories and analysed by two independent operators, were different.

2.3.3 Variable selection

In addition to the different pre-treatments techniques investigated, a variable selection procedure was performed using interval Partial Least Square-Discriminant Analysis (iPLS-DA) algorithm (Nørgaard et al., 2000), in order to identify the best subset of spectral variables to consider for the elaboration of the PLS-DA classification models. The iPLS-DA algorithm, as well as other interval-based chemometric techniques available in literature (Savorani et al., 2013), consists essentially in subdividing the whole spectral range in equidistant intervals (with arbitrary-defined length) and building a series of 'local' models, first using one interval at a time and then adding the intervals in an iterative way. The best model in terms of selected variables is defined by the lowest Root Mean Square Error value achieved in cross-validation (RMSECV) (Leardi and Nørgaard, 2004). In the present work, four different interval size values (i) were considered, with length of 80, 40, 20 or 10 spectral variables, respectively.

2.3.4 Cross-laboratory model validation

As a final step of data elaboration, a validation of the classification models was performed by predicting the coffee origin, always considering both continent and country, using at the same time the NIR spectra acquired from UNIMI and POLITO laboratories. At first, the best iPLS-DA model developed with the POLITO calibration set was validated on the UNIMI external test set.

Subsequently, the model build using the UNIMI calibration set and the same spectral variables selected for the abovementioned iPLS-DA model was used to predict the POLITO external test set. Weighted average efficiency in prediction (EFF_{TS}) was calculated to assess models' reliability.

3. Results and discussion

3.1 Spectral features

FT-NIR spectra collected in both laboratories are characterised by absorption bands related to different vibrations of the chemical bonds. No matter the country of origin, the main observable absorption bands are related to different organic chemical compounds of coffee, mainly caffeine, lignin and fatty acids (Figure 1a). In particular, absorption associated with the C- H bonds of caffeine can be distinguished between 5600 and 4000 cm^{-1}; in the same area contributes the absorption of C- H bonds of fatty acids, amino acids and lignin, which also show absorption bands around 8300-8200 cm^{-1}. O-H absorptions are observable around 10000 cm^{-1}, 6800 cm^{-1} and 4800 cm^{-1} and are mainly linked to cellulose. Moreover, cellulose signal around 4400 cm^{-1} is ascribable to the C- H and O- H bonds. Absorptions related to N- H bonds of polyamides (6800 cm^{-1}) and proteins (6400, 5700 and 4800 cm^{-1}) are also distinguishable (Buratti et al., 2015; Weyer and Workman, 2007).

Even if the spectral features of coffee beans are the same, few differences exist between the SNV transformed average spectrum of the American samples and the one of the Asian samples (Figure 1b). In particular, the average spectrum of the American samples shows higher absorption for the broad band at 8450-8000 cm^{-1}, for the band at 7000-6500 cm^{-1} and the consecutive shoulder up to 6100 cm^{-1} and again between 6000-5400 cm^{-1}, in agreement with what observed by Marquetti et al., 2016. Conversely, the average spectrum of the Asian samples is characterized by higher absorbance values from 12500 to 10500 cm^{-1}, which can be related to colour effect from the adjacent visible region. Moreover, higher absorbance for these samples can be observed at around 5200 cm^{-1} and for the peak at 4200 cm^{-1}.

Figure 1

3.2 Exploratory data analysis

PCA was applied on all the pre-processed NIR dataset to assess which kind of pre-treatments better separated the coffee samples based on both the continent and the country of origin.

Both the PCA models calculated independently from UNIMI and POLITO laboratories led to a rather satisfactory discrimination after the combined application of SNV and mean centering on the raw NIR spectra. The plots reported in Figure 2 refer to the spectra collected from POLITO. It is noteworthy that the same clustering trends shown in Figure 2 can be observed also for the UNIMI data (Figure 1S). The scores plot of the first two Principal Components (Figure 2a) shows that PC1 (82.61% of captured variance), is the main direction along which the samples separated according to the continent of origin. The greatest amount of American coffee samples assumed negative score values of PC1; thus, they were found in the left side of the scores plot. Conversely, almost all the Asian coffee samples were found to assume positive score values of PC1, thus they were gathered in the right side of the score plot. The observed score distribution is in accordance with Medina et al. (2017), who identified a clear separation between Arabica (American coffee beans) and Robusta (Indian coffee beans) species using NIR spectroscopy. However, in our study, some of the Asian coffee samples appear overlapped to the American ones, at least looking at the score plot of the first two components.

The distribution of the scores along PC2 (12.69% of explained variance) could be essentially ascribed to the potential intrinsic variability existing among coffee samples belonging to the same continent but to different countries. In particular, a great variability was found among the Asian samples, which even formed two well-defined clusters. The first one, which included the greatest part of the samples, is characterized by positive score values on PC2, while the second cluster is characterized by negative score values on the same component.

Figure 2b represents the score plot (PC1 vs. PC2) of the same PCA model presented in Figure 2a, but where each coffee sample was labelled according to the respective country of origin. As before, a separation trend can be found mainly along PC1. The Honduran samples assumed the lower score values, followed by Brazilian samples with score values approximately near to zero. The majority of Indian samples, as well as the Vietnamese samples, assumed instead positive score values, but no trend of separation between these two countries was observed along PC1, which was instead found along PC2. Moreover, some Indian samples appeared to be quite overlapped both to Honduran and Brazilian samples. This is in accordance with the findings of Bona et al., 2017, who tested within-country samples and found a clear grouping of coffee beans according to the considered Brazilian cities along PC1 (84% of variance) but a relevant group scattering along PC2 (12% of variance), probably related to the genetic variability and the different cultivation conditions.

The samples belonging to the class 'Other' were found to be spread over all the plot quarters without grouping in a defined cluster, neither along PC1 nor along PC2. This could be expected as they belong to different countries located in different continents. Indeed, the 7 samples located in the PC1 positive quarter have Indonesian origin, whereas the samples from Guatemala (11), Colombia (8), Costa Rica (2) and Nicaragua (2) assumed negative PC1 values except for one sample. Due to the lack of homogenous distribution of this class, its implementation in the further classification models could add errors more than benefits; thus, it was excluded.

To relate the samples distribution observed in the scores plot to spectral features, the loadings plot of PC1 and PC2 are represented in Figure 2c. It was found that at many spectral regions commented in Section 3.1 the corresponding loading values are different from zero. This means that these regions contribute in explaining the variability of the coffee samples on both the PCs considered. By looking at the loading values on PC1 it is possible to state that negative loadings values are associated with spectral regions where the NIR signal is higher for the American samples, indeed having negative score values on PC1; whereas the spectral regions with positive loading values denote a higher NIR signal for the Asian samples; actually they have positive score values on PC1. In detail, variables negatively influencing the samples distribution along PC1 are the ones corresponding to the maximum of broad peak at 8450-8000 cm^{-1}, the beginning of the peak at 7000 cm^{-1} and the two small peaks at 5780 and 5680 cm^{-1}. The variable positively influencing samples distribution along PC1 is the one corresponding to 4020 cm^{-1}. Concerning PC2 loading values, spectral variables at 5200 and 4020 cm^{-1} pull samples to negative values; whereas the change in spectral slope at 7000 cm^{-1} and the peaks at 6000 and 5400 cm^{-1} influenced samples distribution towards positive PC1 score values.

Figure 2

3.3 PLS-DA classification models

The continent-based classification models were built for the spectra collected in each laboratory after subjecting them to all the considered pre-treatment techniques.

The models showed a good classification capability, leading in any case to EFF values in prediction of the external test set (EFF_{TS}) higher than 93.0% (Table 2). Moreover, the model robustness and reproducibility were proved by the high similarity of the results in calibration, cross-validation and validation of the external test set, no matter the pre-treatment applied. The model performances agree with those reported by Medina et al., 2017 for the PLS-DA classification of coffee beans by species (Arabica vs. Robusta), which led to 100% (model) accuracy using a 7-fold cross-validation.

However, they did not perform an external validation as it was done in the present work for making the results even more robust. Indeed, the results of the McNemar test demonstrated that all the models obtained, no matter the pre-treatment or the laboratory performing the analysis, were comparable in terms of classification error rate in prediction ($P > 0.05$, data not shown). Thus, the best PLS-DA classification results could be considered the ones obtained after SNV transformation for both laboratories, since they combine a soft mathematical data pre-treatment and a high EFF in prediction, giving results analogous to those calculated with any other pre-treated data.

Table 2

Regarding the country-based classification models (Table 3), the highest performance in cross-validation was obtained for the Brazilian samples, with 98.1-95.9% of EFF_{CV} for all the models calculated, except for those developed after second derivative transformation. Honduras, India and Vietnam classes were better predicted when soft pre-processing strategies (namely SNV and MSC) were applied, leading to an EFF_{CV} ranging from 97.8 to 90.0%. The results obtained from the prediction of the external test set confirmed what already observed for the internal validation of the model. The PLS-DA results reported by Marquetti et al., 2016 showed a better performance, reaching up to 100 % of sensitivity and specificity. Nevertheless, it should be mentioned that, in that case, the model was validated using 18 samples whereas, in the present study, 47 samples belonging to two deliveries, were considered as validation set.

Table 3

To further assess the models' robustness, the classification performance in prediction was evaluated by McNemar test giving an objective comparison of the results.

From the comparison of the results developed with the FT-NIR data collected at POLITO, it resulted that all the models gave comparable performances ($P > 0.05$), except the ones obtained after second derivative transformation alone or combined with SNV and MSC (Table 4).

The best model developed with FT-NIR spectra collected at UNIMI was the one combining SNV and first derivative transformation, which led to EFF_{TS} of 98.5, 98.7, 93.5 and 83.7% for Brazil, Honduras, India and Vietnam, respectively (Table 3). However, the McNemar results demonstrated that this model has no significant differences ($P > 0.05$) if compared with the performances in prediction of raw, SNV and MSC models always developed with the data collected in the UNIMI laboratory (Table 4).

Comparing POLITO and UNIMI results by McNemar test, it was observed that the models developed with no pre-treatment (raw data) or with SNV, SNV combined with first derivative, and MSC pre-processing gave comparable prediction performances ($P > 0.05$, Table 4).

Table 4

3.4 Variable selection

In order to improve the model performances and in vision of a simplified instrument with reduced spectral range, a variable selection strategy (iPLS-DA) was applied for the best models obtained with the whole spectral range.

Since a soft mathematical pre-treatment could be considered the better solution for practical implementation (Grassi et al., 2018a), the models developed with SNV transformed data could be considered the most convenient strategy to be further optimized by variable selection for both laboratories.

Therefore, the continent-based models developed with SNV data by both laboratories were subjected to iPLS-DA variable selection, as reported in Section 2.3.3. The EFF_{TS} of the iPLS-DA models was always 100% for America-class and 96.5% for Asia-class, no matter the data points interval size (*i*-80, *i*-40, *i*-20 or *i*-10) (data not shown). These results are in agreement with the variable reduction approach proposed by Calvini et al., 2017, who obtained classification models with EFF_{TS} of 98.3% and 100.0% for Arabica and Robusta, respectively, by applying different filter simulations on NIR-hyperspectral data.

Moreover, the McNemar test was performed to compare the results obtained with the whole spectra range and after the variable selection. P-values higher than 0.05 confirmed the rejection of the null hypothesis, meaning that no significant difference exists between the models. Thus, a model developed with just 40 spectral variables out of 1154 (i.e. the whole spectral range) selected in four different spectral regions (12258-12188 cm^{-1}, 5855-5786 cm^{-1}, 5315-5246 cm^{-1} and 4852-4783 cm^{-1}) was found to ensure an efficient continent-based discrimination, decreasing considerably the NIR instrument complexity at the same time.

In the case of country-based classification, the iPLS-DA models gave EFF_{TS} ranging from 100% to 94.9%, depending on the class and the interval size considered (Table 5).

Table 5

The excellent performances of the country-based models were confirmed by McNemar results that reported no significant differences ($P > 0.05$) among all the iPLS-DA models considered. Thus, in

this case, 90 spectral variables selected in three different spectral regions (9018-8871 cm^{-1}, 8632-8177 cm^{-1} and 6009-5940 cm^{-1}) guarantee discriminant performances comparable to the whole spectral range SNV-models. This is in accordance with the use of portable devices with spectral range reduced from 12500-4000 cm^{-1} to 10500-6000 cm^{-1} and only 125 variables, which demonstrated to give reliable results for both food authentication (Grassi et al, 2018a) and process monitoring (Grassi et al., 2018b).

3.5 Cross-laboratory model validation

The iPLS-DA continent-based classification model developed using the 40 abovementioned spectral variables was validated cross-laboratory. Both combinations, i.e. calibration set by POLITO and external set by UNIMI and vice versa, gave a weighted average EFF_{TS} of 100%

Concerning the country-based discrimination, the iPLS-DA model developed with the POLITO training set gave a weighted EFF_{TS} of 95.9% when validated on the UNIMI external test set, while the prediction of the POLITO external test set from the iPLS-DA model calibrated on the UNIMI training dataset led to a weighted EFF_{TS} equal to 94.6%. Therefore, in all cases, the prediction performances were comparable to those achieved by the iPLS-DA classification models calibrated and validated using the NIR spectra acquired by the same laboratory.

4. Conclusions

In the present study, the feasibility of implementing an automated system for the determination of the geographical origin of green coffee beans, based on NIR spectroscopy and multivariate data analysis, has been demonstrated.

Besides the automation, the main benefits related to this kind of analytical approach are the objectivity, the non-destructive nature and its rapidity, leading to a cost-effective improvement of the quality assurance of such a key worldwide food product. Actually, with appropriate industrialization and once the chemometric model has been properly calibrated, the time elapsed from the acquisition of NIR spectra on unknown samples and their subsequent classification would require just a few seconds. Therefore, this method could represent a concrete and effective answer to the need, claimed by coffee producers, industrial manufacturers, as well as by the Food Control Authority, of affordable, rapid and efficient technologies for the evaluation of food quality and authenticity (in this case applied to coffee beans).

Moreover, the variable selection results establish the groundwork for the development of a portable and cost-effective handheld NIR device, customised for origin discrimination of green coffee beans directly "in-field" to certify authenticity and counteract frauds.

Last but not least, the promising results achieved by the cross-laboratory model validation demonstrate the potential transferability of a NIR spectroscopy-based method among different production sites or industries, where the availability of more instruments and different operators is required to perform routine quality control analyses.

References

- Alessandrini, L., Romani. S., Pinnavaia, G., & Dalla Rosa, M. (2008). Near infrared spectroscopy: An analytical tool to predict coffee roasting degree. *Analytica Chimica Acta*, *625(1)*, 95-102.
- Alonso-Salces, R. M., Serra, F., Reniero, F., & Héberger, K. (2009). Botanical and Geographical Characterization of Green Coffee (*Coffea arabica* and *Coffea canephora*): Chemometric Evaluation of Phenolic and Methylxanthine Contents. *Journal of Agricultural and Food Chemistry*, *57(10)*, 4224-4235.
- Anderson, K. A., & Smith, B. W. (2002). Chemical Profiling to Differentiate Geographic Growing Origins of Coffee. *Journal of Agricultural and Food Chemistry*, *50(7)*, 2068-2075.
- Arana, V. A., Medina, J., Alarcon, R., Moreno, E., Heintz, L., Schäfer, H., & Wist, J. (2015). Coffee's country of origin determined by NMR: The Colombian case. *Food Chemistry*, *175*, 500-506.
- Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods, 5,* 3790-3798.
- Bertone, E., Venturello, A., Giraudo, A., Pellegrino, G., & Geobaldo, F. (2016). Simultaneous determination by NIR spectroscopy of the roasting degree and Arabica/Robusta ratio in roasted and ground coffee. *Food Control*, *59*, 683-689.
- Bicchi, C. P., Panero, O. M., Pellegrino, G. M., & Vanni, A. C. (1997). Characterization of roasted coffee and coffee beverages by solid phase microextraction-gas chromatography and principal component analysis. *Journal of Agricultural and Food Chemistry*, *45(12)*, 4680-4686.
- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods, 6,* 2812-2831.

- Boix, A., Fernández Pierna, J. A., von Holst, C., & Baeten, V. (2012). Validation of a near infrared microscopy method for the detection of animal products in feedingstuffs: results of a collaborative study. *Food Additives & Contaminants: Part A, 29(12),* 1872-1880.
- Bona, E., Marquetti, I., Link, J. V., Makimori, G. Y. F., da Costa Arca, V., Guimarães Lemes, A. L., Ferreira, J. M. G., dos Santos Scholz, M. B., Valderrama, P., & Poppi, R. J. (2017). Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee. *LWT-Food Science and Technology*, *76*, 330-336.
- Buratti, S., Sinelli, N., Bertone, E., Venturello, A., Casiraghi, E., & Geobaldo, F. (2015). Discrimination between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy, electronic nose and electronic tongue analysis. *Journal of the Science of Food and Agriculture*, *95(11)*, 2192-2200.
- Cagliani, L. R., Pellegrino, G., Giugno, G., & Consonni, R. (2013). Quantification of *Coffea arabica* and *Coffea canephora* var. Robusta in roasted and ground coffee blends. *Talanta*, *106*, 169-173.
- Calvini, R., Amigo, J. M., & Ulrici, A. (2017). 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*, *967*, 33-41.
- Carrera, F., Leon-Camacho, M., Pablos, F., & Gonzalez, A. G. (1998). Authentication of green coffee varieties according to their sterolic profile. *Analytica Chimica Acta*, *370*, 131-139.
- Casal, S., Alves, M. R., Mendes, E., Oliveira, M. B. P. P., & Ferreira, M. A. (2003). Discrimination between Arabica and Robusta coffee species on the basis of their amino acid enantiomers. *Journal of Agricultural and Food Chemistry*, *51(22)*, 6495-6501.
- El-Abassy, R. M., Donfack, P., & Materny, A. (2011). Discrimination between Arabica and Robusta green coffee using visible micro Raman spectroscopy and chemometric analysis. *Food Chemistry*, *126(3)*, 1443-1448.
- Esteban-Díez, I., González-Sáiz, J. M., & Pizarro, C. (2004). Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy. *Analytica Chimica Acta*, *525(2)*, 171-182.
- Esteban-Díez, I., González-Sáiz, J. M., Sáenz-González, C., & Pizarro, C. (2007). Coffee varietal differentiation based on near infrared spectroscopy. *Talanta, 71(1)*, 221-229.
- Fagerlan, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, *13*, 1-8.

- Frega, N., Bocci, F., & Lercker, G. (1995). Determination of Robusta in commercial coffee blends with Arabica coffee. *Industrie Alimentari*, *34(339)*, 705-708.
- Gil-Agusti, M. T., Campostrini, N., Zolla, L., Ciambella, C., Invernizzi, C., & Righetti, P. G. (2005). Two-dimensional mapping as a tool for classification of green coffee bean species. *Proteomics*, *5(3)*, 710-718.
- González, A. G., Pablos, F., Martín, M. J., León-Camacho, M., & Valdenebro, M. S. (2001). HPLC analysis of tocopherols and triglycerides in coffee and their use as authentication parameters. *Food Chemistry*, *73(1)*, 93-101.
- Grassi, S., Casiraghi, E., & Alamprese, C. (2018a). Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chemistry*, *243*, 382-388.
- Grassi, S., Cardone, G., Bigagnoli, D., & Marti, A. (2018b). Monitoring the sprouting process of wheat by non-conventional approaches. *Journal of Cereal Science*, *83*, 180-187.
- Huck, C. W., Guggenbichler, W., & Bonn, G. K. (2005). Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy (NIRS) compared to high-performance liquid chromatography (HPLC) coupled to mass spectrometry. *Analytica Chimica Acta*, *538*, 195-203.
- Kemsley, E. K., Ruault, S., & Wilson, R. H. (1995). Discrimination between *Coffea arabica* and *Coffea canephora* variant Robusta beans using infrared-spectroscopy. *Food Chemistry*, *54(3)*, 321-326.
- Leardi, R., & Nørgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics, 18(11),* 486-497.
- Marquetti, I., Link, J. V., Guimarães Lemes, A. L., dos Santos Scholz, M. B., Valderrama, P., & Bona, E. (2016). Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of Arabica coffee. *Computers and Electronics in Agriculture*, *121*, 313-319.
- Martín, M. J., Pablos, F., González, A. G., Valdenebro, M. S., & León-Camacho, M. (2001). Fatty acid profiles as discriminant parameters for coffee varieties differentiation. *Talanta*, *54(2)*, 291-297.
- Medina, J., Caro Rodríguez, D., Arana, V. A., Bernal, A., Esseiva, P., & Wist, J. (2017). Comparison of Attenuated Total Reflectance Mid-Infrared, Near Infrared, and 1H-Nuclear Magnetic Resonance Spectroscopies for the Determination of Coffee's Geographical Origin. *International Journal of Analytical Chemistry*. https://doi.org/10.1155/2017/7210463.

- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy, 54 (3),* 413-419.
- Ozaki, Y., McClure W. F., & Christy, A. A. (2006). *Near-Infrared Spectroscopy in Food Science and Technology*. United States of America: Wiley & Sons, Inc.
- Patz, C. D., Blieke, A., Ristow, R., & Dietrich, H. (2004). Application of FT-MIR spectrometry in wine analysis. *Analytica Chimica Acta*, *513(1)*, 81-89.
- Pizarro, C., Esteban-Díez, I., González-Sáiz, J. M., & Forina, M. (2007). Use of near-infrared spectroscopy and feature selection techniques for predicting the caffeine content and roasting color in roasted coffees. *Journal of Agricultural and Food Chemistry*, *55 (18)*, 7477-7488.
- Ribeiro, J. S., Ferreira, M. M. C., & Salva, T. J. G. (2011). Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy. *Talanta*, *83(5)*, 1352-1358.
- Roggo, Y., Duponchel, L., & Huvenne, J. P. (2003). Comparison of supervised pattern recognition methods with McNemar's statistical test: Application to qualitative analysis of sugar beet by near-infrared spectroscopy. *Analytica Chimica Acta*, *477(2)*, 187-200.
- Rubayiza, A. B., & Meurens, M. (2005). Chemical discrimination of Arabica and Robusta coffees by Fourier transform Raman spectroscopy. *Journal of Agricultural and Food Chemistry*, *53(12)*, 4654-4659.
- Savorani, F., Rasmussen, M. A., Rinnan, Å., & Engelsen, S. B. (2013). Chapter 12 - Interval-based chemometric methods in NMR foodomics. *Data Handling in Science and Technology, 28,* 449-486.
- Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8,* 3-16.
- Teuber, R. (2010). Geographical Indications of Origin as a Tool of Product Differentiation: The Case of Coffee. *Journal of International Food & Agribusiness Marketing*, *22 (3-4)*, 277-298.
- Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In: *A Practical Approach to Microarray Data Analysis* (pp. 91-109). Boston: Springer.

- Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *Journal of Scientific and Statistical Computing*, *5 (3)*, 735-743.
- Weyer, L., & Workman Jr, J. (2007). *Practical guide to interpretive near-infrared spectroscopy*. Boca Raton: CRC press.

563 **Figure captions**

564 **Figure 1 -** Average NIR spectrum of American coffee samples (blue line) and Asian coffee samples
565 (red line): **(a)** from raw data and **(b)** after SNV pre-treatment. (For interpretation of the references
566 to colour in this figure legend, the reader is referred to the web version of this article.)

567 **Figure 2 -** PCA results of spectra collected by POLITO and pre-treated by SNV and mean
568 centering: **(a)** Scores (PC1 vs. PC2) of coffee samples labelled according to the continent of origin;
569 **(b)** scores (PC1 vs. PC2) of coffee samples labelled according to the country of origin; **(c)** loadings
570 plot for PC1 and PC2. (For interpretation of the references to colour in this figure legend, the reader
571 is referred to the web version of this article.)

572 **Figure 1S -** PCA results of spectra collected by UNIMI and pre-treated by SNV and mean
573 centering: **(a)** Scores (PC1 vs. PC2) of coffee samples labelled according to the continent of origin;
574 **(b)** scores (PC1 vs. PC2) of coffee samples labelled according to the country of origin; **(c)** loadings
575 plot for PC1 and PC2. (For interpretation of the references to colour in this figure legend, the reader
576 is referred to the web version of this article.)

577

Highlights

- NIR Spectroscopy investigated to identify the geographical origin of green coffee
- Partial Least Square-Discriminant Analysis applied to build classification models
- Coffee origin for both continent and country was tested as discrimination parameter
- Variable selection applied to NIR spectra allowed improving the model performance
- Inter-laboratory comparison of the classification results was made by McNemar test

Table 1. Continents and countries of origin of green coffee samples.

Continent	*Country*	*Sample no.*
AMERICA	BRAZIL	38
	HONDURAS	27
	GUATEMALA	11
	COLOMBIA	8
	COSTA RICA	2
	NICARAGUA	2
ASIA	INDIA	65
	VIETNAM	31
	INDONESIA	7

Table 2. Results of continent-based Partial Least Squares-Discriminant Analysis (PLS-DA) on NIR spectra of green coffee samples after different mathematical pre-treatments: Efficiency percentage obtained in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS).

		AMERICA			ASIA		
		CAL	***CV***	***TS***	***CAL***	***CV***	***TS***
POLITO	Raw	98.6	98.6	100.0	94.6	94.6	100.0
	SNV	98.6	98.6	100.0	93.2	93.2	93.1
	SNV + d1	98.6	98.6	100.0	93.2	93.2	96.6
	SNV + d2	98.6	98.6	100.0	95.9	93.2	96.6
	MSC	98.6	98.6	100.0	93.2	93.2	93.1
	MSC + d1	98.6	98.6	100.0	93.2	93.2	96.6
	MSC + d2	98.6	98.6	100.0	95.9	93.2	96.6
	d1	98.6	98.6	100.0	95.9	95.9	100.0
	d2	98.6	98.6	100.0	97.3	94.6	96.6
UNIMI	Raw	98.6	98.6	100.0	97.3	97.3	100.0
	SNV	98.6	98.6	100.0	93.2	91.9	96.5
	SNV + d1	98.6	98.6	100.0	94.6	93.2	93.1
	SNV + d2	97.2	97.2	100.0	94.6	93.2	96.5
	MSC	98.6	98.6	100.0	93.2	91.9	96.5
	MSC + d1	98.6	98.6	100.0	94.6	93.2	93.1
	MSC + d2	97.2	97.2	100.0	94.6	93.2	96.5
	d1	98.6	98.6	100.0	91.9	91.9	96.5
	d2	97.2	97.2	100.0	95.9	93.2	96.5

Note. POLITO: models developed from the NIR data collected at the Department of Applied Science and Technology (Polytechnic of Turin); UNIMI: models developed from the NIR data collected at the Department of Food, Environmental and Nutritional Sciences (University of Milan). Raw: data elaborated without any pre-treatment except mean centering; SNV: standard normal variate; MSC: multiplicative scatter correction; d1: first derivative; d2: second derivative.

Table 3

Table 3. Results of country-based Partial Least Squares-Discriminant Analysis (PLS-DA) on NIR spectra of green coffee samples after different mathematical pre-treatments: Efficiency percentage obtained in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS).

		BRAZIL			HONDURAS			INDIA			VIETNAM		
		CAL	*CV*	*TS*	*CAL*	*CV*	*TS*	*CAL*	*CV*	*TS*	*CAL*	*CV*	*TS*
POLITO	Raw	98.1	97.6	100.0	97.3	92.0	100.0	94.1	92.3	98.1	95.1	92.1	94.9
	SNV	97.6	97.6	100.0	99.5	95.5	100.0	97.5	91.6	93.5	95.1	92.1	88.2
	SNV + d1	97.6	97.6	98.5	95.0	95.0	98.7	93.4	92.3	89.0	92.6	92.1	83.7
	SNV + d2	97.0	97.0	98.5	92.8	91.0	93.5	86.5	86.4	76.1	92.1	89.5	77.5
	MSC	97.6	97.6	100.0	99.5	95.5	100.0	97.5	92.3	93.5	95.1	94.6	88.2
	MSC + d1	97.6	97.6	98.5	95.0	95.0	98.7	93.4	92.3	89.0	92.6	92.1	83.7
	MSC + d2	97.0	97.0	98.5	92.8	91.0	93.5	86.5	86.4	76.1	92.1	89.5	77.5
	d1	97.6	97.0	100.0	95.0	92.6	98.7	93.4	90.9	89.0	92.6	86.8	82.5
	d2	95.9	90.4	97.0	91.0	85.6	94.9	84.0	82.1	76.1	92.6	89.5	77.5
UNIMI	Raw	97.6	97.6	100.0	96.8	94.0	100.0	90.5	89.4	92.0	90.0	89.5	77.5
	SNV	98.1	98.1	100.0	98.4	97.8	100.0	95.0	93.9	92.0	90.0	90.0	77.5
	SNV + d1	98.1	96.5	98.5	98.4	90.4	98.7	95.0	93.9	93.5	90.0	87.3	83.7
	SNV + d2	97.6	94.0	98.5	92.9	86.5	93.5	84.6	85.8	74.3	86.8	86.4	70.7
	MSC	98.1	98.1	100.0	98.4	97.8	100.0	95.0	93.9	92.0	90.0	90.0	77.5
	MSC + d1	97.6	95.9	98.5	92.9	83.6	93.5	85.8	85.8	74.3	87.3	86.8	70.7
	MSC + d2	97.6	94.0	98.5	92.9	86.0	93.5	84.6	84.6	74.3	86.8	86.4	70.7
	d1	97.0	96.5	98.5	90.7	86.5	93.5	85.1	86.2	74.3	84.5	84.1	70.7
	d2	95.1	94.0	98.5	90.1	86.0	94.9	84.6	86.9	74.3	86.8	87.3	69.7

Note. POLITO: models developed from the NIR data collected at the Department of Applied Science and Technology (Polytechnic of Turin); UNIMI: models developed from the NIR data collected at the Department of Food, Environmental and Nutritional Sciences (University of Milan). Raw: data elaborated without any pre-treatment except mean centering; SNV: standard normal variate; MSC: multiplicative scatter correction; d1: first derivative; d2: second derivative.

Table 4

Table 4. Results of McNemar test: P-value resulting from pair comparison of the country-based Partial Least Squares-Discriminant Analysis (PLS-DA) on NIR spectra of green coffee samples after different mathematical pre-treatments. In bold, the p-values <0.05 denote a significant difference between the considered models.

		POLITO									**UNIMI**								
		Raw	SNV	SNV + d1	SNV + d2	MSC	MSC + d1	MSC + d2	d1	d2	Raw	SNV	SNV + d1	SNV + d2	MSC	MSC + d1	MSC + d2	d1	d2
	Raw	1.000	0.250	0.063	0.002	0.250	0.063	**0.002**	0.063	**0.002**	0.125	0.125	0.125	**0.001**	0.125	**0.001**	**0.001**	**0.001**	**0.001**
	SNV	0.250	1.000	0.250	**0.021**	1.000	0.250	**0.021**	0.250	**0.021**	0.625	0.625	0.500	**0.012**	0.625	**0.012**	**0.012**	**0.012**	**0.012**
	SNV + d1	0.063	0.250	1.000	0.070	0.250	1.000	0.070	1.000	0.070	0.625	0.688	0.625	**0.039**	0.688	**0.039**	**0.039**	**0.039**	**0.039**
	SNV + d2	**0.002**	**0.021**	0.070	1.000	**0.021**	0.070	1.000	0.070	1.000	**0.039**	**0.039**	0.065	0.500	**0.039**	0.500	0.500	0.500	0.500
POLITO	MSC	0.250	1.000	0.250	**0.021**	1.000	0.250	**0.021**	0.250	**0.021**	0.625	0.625	0.500	**0.012**	0.625	**0.012**	**0.012**	**0.012**	**0.012**
	MSC + d1	0.063	0.250	1.000	0.070	0.250	1.000	0.070	1.000	0.070	0.625	0.688	0.625	**0.039**	0.688	**0.039**	**0.039**	**0.039**	**0.039**
	MSC + d2	**0.002**	0.021	0.070	1.000	0.021	0.070	1.000	0.070	1.000	**0.039**	**0.039**	0.065	0.500	**0.039**	0.500	0.500	0.500	0.500
	d1	0.063	0.250	1.000	0.070	0.250	1.000	0.070	1.000	0.070	0.625	0.688	0.625	**0.039**	0.688	**0.039**	**0.039**	**0.039**	**0.039**
	d2	**0.002**	**0.021**	0.070	1.000	**0.021**	0.070	1.000	0.070	1.000	**0.039**	**0.039**	0.065	0.500	**0.039**	0.500	0.500	0.500	0.500
	Raw	0.125	0.625	0.625	**0.039**	0.625	0.625	0.039	0.625	**0.039**	1.000	1.000	1.000	**0.008**	1.000	**0.008**	**0.008**	**0.008**	**0.008**
	SNV	0.125	0.625	0.688	**0.039**	0.625	0.688	0.039	0.688	**0.039**	1.000	1.000	1.000	**0.008**	1.000	**0.008**	**0.008**	**0.008**	**0.008**
	SNV + d1	0.125	0.500	0.625	0.065	0.500	0.625	0.065	0.625	0.065	1.000	1.000	1.000	**0.021**	1.000	**0.021**	**0.021**	**0.021**	**0.021**
	SNV + d2	**0.001**	**0.012**	**0.039**	0.500	**0.012**	**0.039**	0.500	**0.039**	0.500	**0.008**	**0.008**	**0.021**	1.000	**0.008**	1.000	1.000	1.000	1.000
UNIMI	MSC	0.125	0.625	0.688	0.039	0.625	0.688	0.039	0.688	**0.039**	1.000	1.000	1.000	0.008	1.000	**0.008**	**0.008**	**0.008**	**0.008**
	MSC + d1	**0.001**	**0.012**	**0.039**	0.500	**0.012**	**0.039**	0.500	**0.039**	0.500	**0.008**	**0.008**	**0.021**	1.000	**0.008**	1.000	1.000	1.000	1.000
	MSC + d2	**0.001**	**0.012**	**0.039**	0.500	**0.012**	**0.039**	0.500	**0.039**	0.500	**0.008**	**0.008**	**0.021**	1.000	**0.008**	1.000	1.000	1.000	1.000
	d1	**0.001**	**0.012**	**0.039**	0.500	**0.012**	**0.039**	0.500	**0.039**	0.500	**0.008**	**0.008**	**0.021**	1.000	**0.008**	1.000	1.000	1.000	1.000
	d2	**0.001**	**0.012**	**0.039**	0.500	**0.012**	**0.039**	0.500	**0.039**	0.500	**0.008**	**0.008**	**0.021**	1.000	**0.008**	1.000	1.000	1.000	1.000

Note. POLITO: models developed from the NIR data collected in the Department of Applied Science and Technology (Polytechnic of Turin); UNIMI: models developed from the NIR data collected in the Department of Food, Environmental and Nutritional Sciences (University of Milan); Raw: data elaborated without any pre-treatment except mean centering; SNV: standard normal variate; MSC: multiplicative scatter correction; d1: first derivative; d2: second derivative.

Table 5

Table 5. Results of iPLS-DA variable selection strategy applied on the country-based Partial Least Squares-Discriminant Analysis (PLS-DA) on the SNV-NIR spectra of green coffee samples: Efficiency percentage obtained in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS).

		BRAZIL			HONDURAS			INDIA			VIETNAM		
		CAL	***CV***	***TS***	***CAL***	***CV***	***TS***	***CAL***	***CV***	***TS***	***CAL***	***CV***	***TS***
POLITO	*i*-80	98.1	98.1	100.0	99.5	98.9	100.0	100.0	97.5	100.0	100.0	95.1	100.0
	i-40	97.6	97.6	100.0	99.5	98.9	100.0	98.9	96.4	100.0	100.0	95.1	100.0
	i-20	98.1	97.6	100.0	99.5	98.4	100.0	100.0	96.0	98.1	100.0	97.6	94.9
	i-10	98.1	98.1	100.0	98.4	97.8	100.0	97.8	95.3	100.0	100.0	95.1	100.0
UNIMI	*i*-80	98.1	98.1	100.0	97.8	97.8	100.0	94.9	94.2	97.3	97.1	94.6	98.6
	i-40	98.1	97.6	100.0	98.9	96.1	100.0	98.2	96.4	98.1	97.6	95.1	94.9
	i-20	97.6	97.0	100.0	98.9	98.9	100.0	96.7	94.2	100.0	99.5	94.6	100.0
	i-10	97.6	97.6	100.0	98.9	98.9	100.0	96.7	95.3	100.0	99.5	94.6	100.0

Note. POLITO: models developed from the NIR data collected in the Department of Applied Science and Technology (Polytechnic of Turin); UNIMI: models developed from the NIR data collected in the Department of Food, Environmental and Nutritional Sciences (University of Milan).

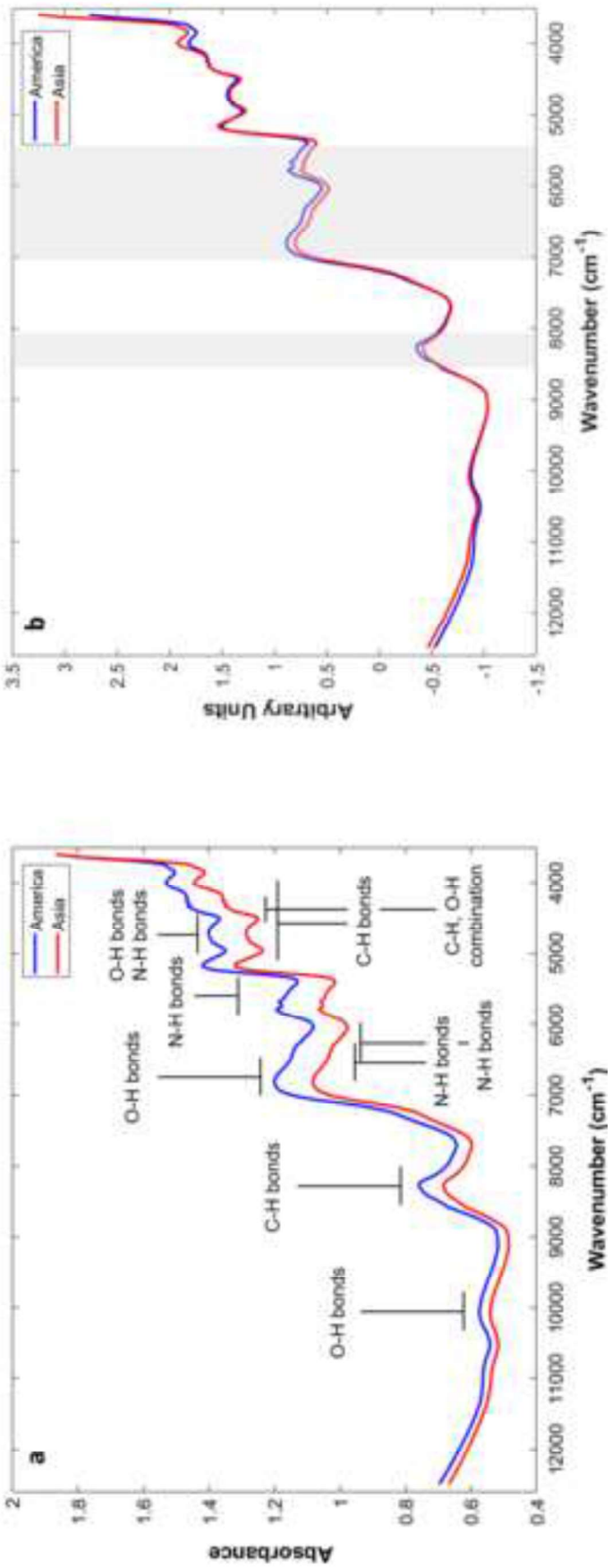
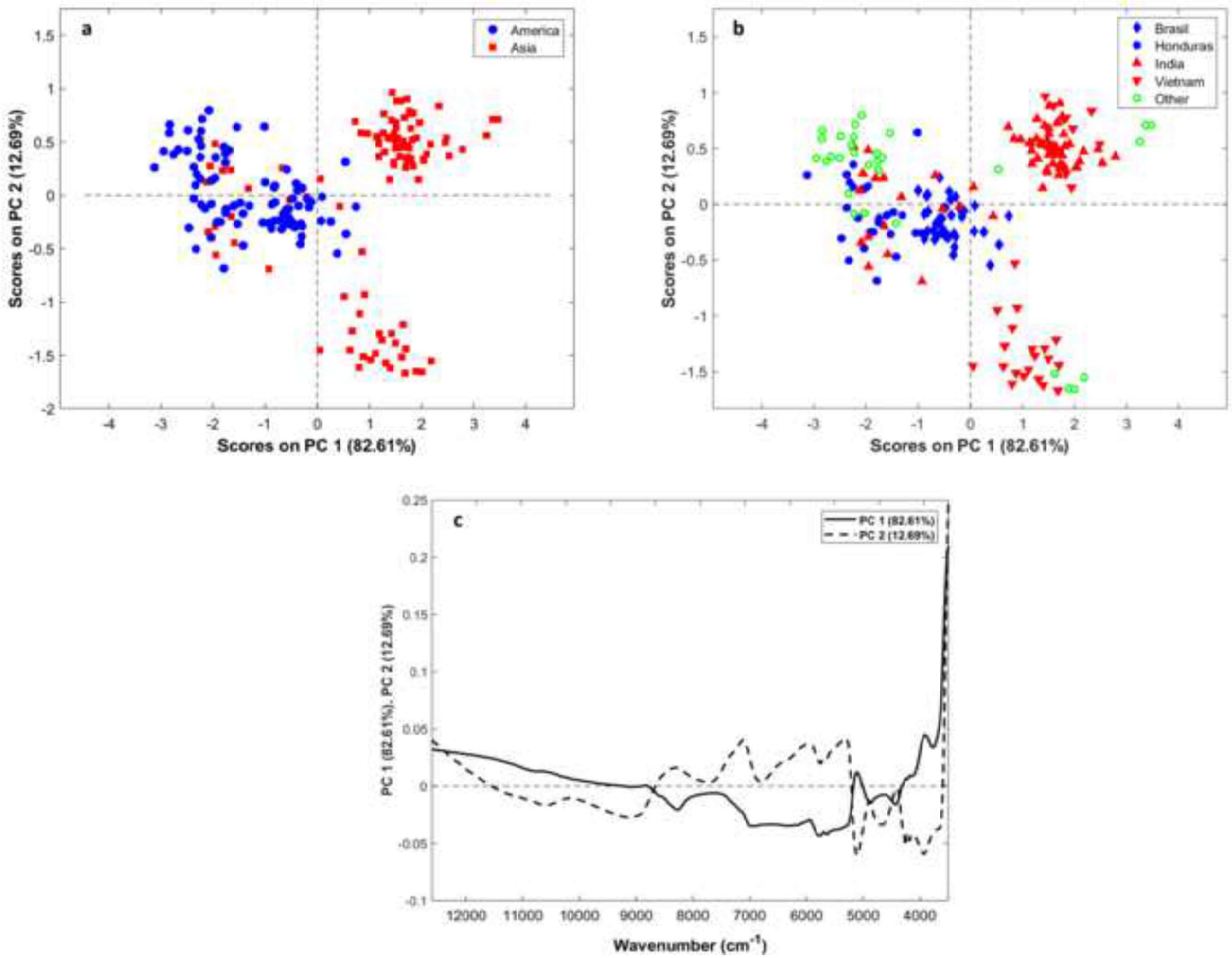


Figure 1
Click here to download high resolution image

Figure 2

Click here to download high resolution image



Supplementary Figure 1

Click here to download high resolution image