Computers and Electronics in Agriculture 121 (2016) 313–319

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee

Izabele Marquetti [a], Jade Varaschim Link [a,b], André Luis Guimarães Lemes [a], Maria Brígida dos Santos Scholz [c], Patrícia Valderrama [a], Evandro Bona [a,*]

[a] Federal University of Technology – Paraná (UTFPR), P.O. Box 271, Via Rosalina Maria dos Santos – 1233, CEP 87301-899 Campo Mourão, PR, Brazil
[b] Federal University of Santa Catarina (UFSC), Campus Universitário – Trindade, CEP 88040-900 Florianópolis, SC, Brazil
[c] Agronomic Institute of Paraná (IAPAR), Rodovia Celso Garcia Cid, km 375, CEP 86047-902 Londrina, PR, Brazil

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 10 December 2015
Accepted 28 December 2015
Available online 14 January 2016

Keywords:
Green coffee
Arabica
Growing region
Genotype
Near infrared spectroscopy
PLS-DA

ABSTRACT

The agronomic practices and environmental conditions for coffee cultivation, such as climate, soil type and altitude, promote influence in the final chemical composition of the grain. Furthermore, the genotype directly influences the essential features of the beverage, increasing its aggregate price. Proof of geographic and genotypic origin of the coffee genotype must be done using reliable methods. Thus, near infrared spectroscopy (NIR) was used to analyze different coffee genotypes that were cultivated in Brazil. Due to complexity and quantity of information within the spectra, partial least square discriminant analysis (PLS-DA) were applied to analyze the NIR data. The multiplicative scatter correction (MSC) and the Savitzky–Golay second-derivative were tested as preprocessing techniques to find which one provides an appropriate identification model. The best model achieved correctly identified 94.4% of validation samples for both geographic and genotypic origin. The results demonstrate that NIR spectroscopy provides significant analytical data to be used in tandem with PLS-DA to distinguish green coffee samples geographically and genotypically.

1. Introduction

Coffee is one of the most consumed beverages in the world, and its consumption has increased due to factors such as improvements in the quality of the beverage, better agronomic practices, its association with health benefits, and availability of new products. The quality of the beverage is associated with adequate coffee genotype selection, which improves its flavor (Farah, 2009). Arabica coffee is known for its high quality, with an intense aroma, lower caffeine content, and a less bitter taste, and so it has higher aggregate price (Lashermes and Anthony, 2007).

There are many arabica coffee genotypes available, most of which were obtained by breeding that modified certain features of the coffee. The objective of developing modern genotypes is to obtain grains more adapted to different climatic conditions and soil types, and also more resistant to diseases and pests, increasing productivity and improving quality (Sera, 2001). The interaction between the genetic variability of arabica and the cultivation conditions affects both the chemical composition as the physicochemical characteristics of the coffee beans (Scholz et al., 2011). The coffee quality as a beverage depends on the chemical composition of green coffee (Ribeiro et al., 2011). Highest quality coffees are related to the increase of sucrose, lipids, amino acids, and trigonelline contents, and reduction of chlorogenic acids and caffeine contents, which are responsible for contributing to the bitterness of the coffee (Stalmach et al., 2006). Beans from regions and varieties that are known to produce high quality beverages have a great commercial value (Teuber, 2010). To guarantee to consumers the geographic and genotypic origin of coffee, fast and efficient analytic methods are required.

Different analytical techniques are often employed for coffee analysis, including chromatographic analysis (Novaes et al., 2015), UV–Vis spectroscopy (Souto et al., 2015), nuclear magnetic resonance (Arana et al., 2015). These are slow techniques because they require more time to prepare samples, have high costs, and generate too much residues. To overcome these disadvantages an alternative is employ near infrared spectroscopy (NIR) which is a fast technique that requires minimum sample preparation, do not destroy samples, and allows simultaneous analyses. Because

[*] Corresponding author. Tel./fax: +55 44 3518 1477.
E-mail address: ebona@utfpr.edu.br (E. Bona).

http://dx.doi.org/10.1016/j.compag.2015.12.018
0168-1699/© 2016 Elsevier B.V. All rights reserved.

it is a technique with a high complexity data and a large amount of information due to overtones and combination bands (mainly stretching and bending vibrations, from C═O, C—H, C—N, C—O, N—H, NO_2, and O—H bonds) chemometric methods are required for spectra interpretation (Burns and Ciurczak, 2008). This technique is often used to discriminate blends of arabica and robusta grains, obtaining satisfactory results (Esteban-Díez et al., 2007; Santos et al., 2012; Bertone et al., 2016) or to detect defects or adulteration in coffees (Craig et al., 2015; Winkler-Moser et al., 2016). However, few studies used NIR spectroscopy for discrimination of arabica coffee by geographical and genotype origins. Then, this study aimed to perform a genotypic and geographical segmentation of coffees grown in Brazil by using NIR spectroscopy coupled with partial least squares with discriminant analysis (PLS-DA).

2. Materials and methods

2.1. Coffee samples and near infrared spectra

Four *Coffea arabica* genotypes developed by the Agronomic Institute of Paraná (IAPAR) were evaluated in this research: IPR 99, IPR 105, IPR 106, and IA 59. The IA 59 genotype was released in 1994, originated from the crossing between *Coffea arabica* varieties “Villa Sarchi 971/10” and “Híbrido de Timor 832/2”. As well as IPR 99, is resistant to all kinds of known rust (Sera et al., 2011). The IPR 105 genotype is derived from the Catuaí genotype, while the IPR 106 genotype originated from the Icatu genotype, and both are equally resistant to all rusts at different levels. Among these genotypes, only IA 59 and IPR 99 genotypes are available to farmers (Sera et al., 2010). To correlate the features of IPR 105 and IPR 106 genotypes with IPR 99 and IA 59 (varieties already cultivated) could contribute for the release of new genotypes in the market. Thus, efficient tools to help find these similarities, especially relating to chemical composition, is extremely important.

Ninety samples of coffee, carefully selected, cultivated in four different cities were evaluated: Cornélio Procópio (CP), Paranavaí (PV), Mandaguari (MD), and Londrina (LD), all in Paraná State in south of Brazil. Five sample of each genotype was used per city, except for IA 59 samples cultivated in PV and CP, where ten samples were used for each city. The samples were harvested between 2008 and 2010. After the harvest, the beans were sent to the IAPAR laboratories in Londrina (Paraná–Brazil), where they were put in wooden boxes with a mesh bottom and moved eight times per day until grain moisture of 11–12% was achieved. Then, the samples were processed by removing the husk and the parchment. The green beans were ground (0.5 mm) and stored in a freezer at −18 °C for later analysis.

Green coffee spectra were recorded using a near infrared spectroscopy NIRSystem 5000-M (Foss Tecator AB, Höganäs, Sweden). Measurements were made at room temperature (23 °C) in the wavelength range 1100–2498 nm at 2 nm intervals. The software WinISI III version 1.50e (Foss NIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA) was used to acquire the spectra.

2.2. Software, preprocess and PLS-DA

All chemometric analyses for the NIR spectra were performed in MATLAB R2007b (The MathWorks Inc., Natick, USA).

To reduce variation sources that carry no relevant information during the multivariate calibration model, and considering scatter effects and slope changes between samples, two spectra pretreatments were applied: multiplicative scatter correction (MSC) and the Savitzky–Golay second-derivative (second order polynomial with a seven points window).

MSC corrects multiplicative and additive scatter effects, which are the result of differences in granules size, morphology, and particle orientation. It uses a linear regression of spectral variables versus the average spectrum (Isaksson and Næs, 1988). The second derivative using Savitzky–Golay transformation removes problems due to slope changes between samples (Savitzky and Golay, 1964).

The PLS-DA is a supervised chemometric method applied for classification problems developed from the two-block Partial Least Squares (PLS) algorithms used in multivariate calibration to model the relationship between two matrices (Barker and Rayens, 2003). In PLS-DA, as in PLS, a linear relationship between the dependent variable (Y) and the independent variable (X) is established. This occur based on principal component analysis (PCA) were the matrix X and Y are decomposed into the product of two matrices, scores and loadings (Wold et al., 1987). In the PLS and PLS-DA methods there are a slight rotation on the principal component axis seeking the maximum covariance of X with Y, and now the principal components are called latent variables (LV). The main difference between PLS and PLS-DA is the Y matrix. The Y matrix in PLS contains the values of the property of interest, while in PLS-DA the matrix Y contains information about the samples classes. For each class, the value is assumed to be 0 or 1, depending on whether or not it belongs to the class represented for that column (Barker and Rayens, 2003). The modeling consists of two steps: (1) calibration, where data characteristics are investigated to find a model for their behavior; and (2) validation, where data that did not participate in the calibration step are used to evaluate the model adequacy.

The threshold value for the class separation is based on Bayes' Theorem. The Bayesian threshold assumes that the predicted y values follow a distribution similar to what will be observed for future samples. Using these estimated distributions, a threshold is selected at the point where the two estimated distributions cross; this is the y value at which the number of false positives and false negatives should be minimized for future predictions. Further the model performance can be evaluated using some parameters such as sensitivity and specificity. The sensitivity is the model ability to correctly classify the samples, relating the predicted samples to being in a class with the samples that actually are in this class. The specificity relates the predicted samples to not being in a class with the samples that actually are not in this class (Almeida et al., 2013).

The model accuracy is evaluated using the root mean square error of calibration (RMSEC) and prediction (RMSEP) (Valderrama et al., 2007, 2009):

$$\mathrm{RMSEC} = \sqrt{\frac{\sum(\hat{y}_p - y_r)^2}{nc - v}} \tag{1}$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum(\hat{y}_p - y_r)^2}{nv}} \tag{2}$$

where nc is the number of samples in calibration set and nv is the number of samples in validation set, v is the number of latent variables plus one when mean centered data is used, $\hat{y}_p$ is the predicted value for each sample and y_r is the reference value for the correspondent sample.

In this work PLS-DA models were built to differentiate arabica coffee samples by genotype and growing region. We used 72 calibration samples and 18 prediction samples, all spectra data were mean centered.

3. Results and discussion

Fig. 1 shows the NIR spectra from coffee samples, and the spectra after employing the preprocess. The different harvest location is a complicating factor during the distinction between samples due