LWT - Food Science and Technology 76 (2017) 330–336

Contents lists available at ScienceDirect

LWT - Food Science and Technology

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

# Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee

Evandro Bona [a, *], Izabele Marquetti [a], Jade Varaschim Link [a, b], Gustavo Yasuo Figueiredo Makimori [a], Vinícius da Costa Arca [a], André Luis Guimarães Lemes [a], Juliana Mendes Garcia Ferreira [a], Maria Brígida dos Santos Scholz [c], Patrícia Valderrama [a], Ronei Jesus Poppi [d]

[a] Post-Graduation Program of Food Technology (PPGTA), 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
[d] Institute of Chemistry, University of Campinas (UNICAMP), P.O. Box 6154, CEP 13083-970 Campinas, SP, Brazil

ARTICLE INFO

*Article history:*
Received 10 February 2016
Received in revised form
20 April 2016
Accepted 25 April 2016
Available online 27 April 2016

*Keywords:*
Machine learning
Near infrared
Mid infrared
Genetic algorithm

ABSTRACT

The coffee is an important commodity to Brazil. Species, climate, genotypes, cultivation practices and industrialization are critical to final quality of the beverage. Thus, the development of analytical methods for coffee authentication is important to ensure the origin of the bean. The purpose of this study was to develop a methodology for geographical classification of different genotypes of arabica coffee using infrared spectroscopy and support vector machines (SVM). The spectra were collected in the range of near infrared (NIRS) and mid infrared (FTIR). For the data analysis, a SVM was built using radial basis as kernel function and the one-versus-all multiclass approach. The C and $\gamma$ parameters of SVM were optimized using the genetic algorithm. With the application of the NIRS-SVM approach all test samples were correctly classified with a sensitivity and specificity of 100%, while FTIR-SVM had a slightly lower performance. Therefore, it was possible to confirm that infrared spectroscopy is a fast and effective method for geographic certification with little sample preparation, and without the production of chemical wastes. Furthermore, the SVM can be a chemometric alternative in tandem with infrared spectroscopy for another classification problems.



## 1. Introduction

The coffee is one of the most consumed beverages in the world, and an important commodity to Brazil, which is the largest producer and exporter in the world. From November 2014 to October 2015, 36 million bags of 60 kg of coffee were exported, corresponding to US$ 6.348 billion (CeCafé, 2015). There are two main species of coffee, *Coffea arabica*, also known as arabica coffee, and *Coffea canephora* or robusta coffee (Clarke & Vitzthum, 2001). These species present a very different chemical composition and arabica coffee is known for its high quality beverage, with an intense aroma, lower caffeine content, and a less bitter taste, showing a higher aggregate price (Lashermes & Anthony, 2007; Link, Guimarães Lemes, Marquetti, dos Santos Scholz, & Bona, 2014). Among cultivars of arabica coffee, different levels of quality beverages have been found due to genetic factors and environmental conditions in the place of cultivation. Beans from regions and varieties that are known to produce high quality beverages have a great commercial value (Bertrand et al., 2012; Joët et al., 2010; Teuber, 2010). However, it is essential to prove the geographical and genotypic origin of the cultivar using a reliable method.

Different analytical techniques are often employed for coffee analysis, including chromatographic analysis (Novaes, Oigman, de Souza, Rezende, & de Aquino Neto, 2015), UV–Vis spectroscopy (Souto et al., 2015), nuclear magnetic resonance (Arana et al., 2015), and physicochemical analysis (Borsato, Pina, Spacino, Scholz, & Androcioli Filho, 2011). These are slow techniques because they require more time to prepare samples, have high costs, and

* Corresponding author.
*E-mail address:* ebona@utfpr.edu.br (E. Bona).

http://dx.doi.org/10.1016/j.lwt.2016.04.048
0023-6438/

generate too much residues. To overcome these disadvantages, an alternative is employ infrared spectroscopy, which is a fast technique that requires minimum sample preparation, do not destroy samples, and allows simultaneous analyses (Downey, Briandet, Wilson, & Kemsley, 1997; Karoui, Downey, & Blecker, 2010; Terouzi et al., 2011). However, infrared spectroscopy is a technique with a high complexity data and a large amount of information. Thus, chemometric methods are required for spectra interpretation (Hiri et al., 2015; Roggo et al., 2007).

In previous studies, the efficiency of Fourier transform mid infrared spectroscopy (FTIR) was verified (Link and Lemes et al., 2014), as well as near infrared spectroscopy (NIRS) (Marquetti et al., 2016) for geographical classification of four arabica coffee genotypes. In the first work, the FTIR spectra were treated by principal component analysis (PCA) to dimensionality reduction, and the scores were used as input to an artificial neural network of radial basis functions (RBF). In the later work, the NIRS spectra were analysed using partial least square with discriminant analysis (PLS-DA). In the present study, twenty arabica coffee genotypes were classified and the spectra was collected in both FTIR and NIR. For data analysis, support vector machine (SVM) was evaluated because it is able to model nonlinear relations. SVM is a new type of machine learning based on statistical learning theory that emphasizes machine learning in the case of fewer samples (Bishop, 2006). The structural risk minimization principle derived from statistical learning theory takes this as the foundation, as compared with RBF, giving the SVM prominent advantages. First, the strong theoretical basis provides high generalization capability and avoids overfitting. In second place, the global model is capable of dealing efficiently with high-dimensional input vectors. Third, the solution is sparse and only a subset of training samples contributes to this solution, thereby reducing the workload (Argyri et al., 2013). As a nonlinear method, SVM shows advantages over PLS-DA, the latest tends to shrink the low variance directions, but can actually inflate some high variance directions. This can make the PLS a little unstable. Besides, PLS down weights noisy features, but does not throw them away; therefore a lot of noise can contaminate the predictions. In addition, highly correlated variables will tend to be chosen together, as a result, there may be much redundancy in the set of selected variables. This might indicate that the PLS method is more prone to overfitting than SVM. Overfitting is more likely to take place on high dimensional data, and infrared spectra typically show very high dimension (Liu, Yang, & Deng, 2015).

The actual objective was to develop a methodology for geographical classification of different genotypes of green arabica coffee using infrared spectroscopy in tandem with support vector machines.

## 2. Materials and methods

### 2.1. Coffee samples

This study used about 3 kg of cherry coffee, harvested between 2009 and 2010, of 20 genotypes collected at four locations (totaling 74 samples) in the coffee region of Paraná-Brazil: Paranavaí, Cornélio Procópio, Mandaguari and Londrina. The samples were placed into wooden boxes with a mesh bottom and moved eight times per day until the beans reached 11–12% moisture. After that, the samples were processed (removal of hull and parchment); the coffee beans were frozen with liquid nitrogen, ground in a mill disk (model Perten 3600) with 0.6-mm final particle size and kept at −18 °C. Before analysis, the samples were thawed and retained in a desiccator to even out moisture (Link, Lemes, et al., 2014; Marquetti et al., 2016). More information about climatic conditions, cities position and the number of samples per genotype, year and city, can be obtained in Electronic Supplementary Material (Tables 1S and 2S).

### 2.2. Fourier transform mid infrared spectroscopy (FTIR)

Pellets were prepared by adding about 100 mg of dry KBr (FTIR grade – Sigma–Aldrich) and approximately 1 mg of finely ground sample. The mixture was compressed in a hydraulic press (Bovenau, P15 ST) using a mold (ICL, ICL's Macro/Micro KBr dye) employing about 35 MPa pressure to produce a transparent pellet. Before the analysis of each sample, the FTIR (Shimadzu FTIR – 8300) was programmed to perform a background spectrum of the air, which was used to subtract the influence of air components in the spectrum. After that, the pastille was positioned on the instrument shaft and the spectra were obtained in the range 4000 to 400 $cm^{-1}$. Accumulated scans ($n = 32$) were used to form the final spectrum and five repetitions (pellets) were performed for each sample, totalizing 370 spectra (Link, Lemes, et al., 2014).

After obtaining the spectra, normalization of the spectrum was done to eliminate effects due to minor differences among the weights of the sample used for the preparation of pellets. Next, baseline correction and smoothing of the spectrum using Savitzky-Golay algorithm was performed (Savitzky & Golay, 1964). Only the spectrum region between 1800 and 800 $cm^{-1}$ was evaluated because it contains the most important absorption bands due to carbonyl axial symmetric deformation (esters, aldehydes, and ketones), methylene angular symmetric deformation, and angular and axial symmetric deformations of C–O (esters and alcohols). Therefore, this region contains the fingerprint information for discrimination of different coffee samples (Link, Lemes, et al., 2014).

### 2.3. Near infrared spectroscopy (NIRS)

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. To reduce variation sources that carry no relevant information during the multivariate calibration, and considering scatter effects between samples, the multiplicative scatter correction (MSC) was applied (Marquetti et al., 2016). 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 & Naes, 1988).

### 2.4. Sample selection

The spectra, both NIRS as FTIR, were split in training set (2/3) and test set (1/3) using the Kennard and Stone algorithm (Kennard & Stone, 1969). In detail, Kennard and Stone algorithm aims at selecting the most diverse set among a given set of candidate samples, to be included in the training set, according to a *maximin* criterion. At first, the distances among all pairs of samples are computed and the two most distant samples are selected to be included in the training set. Successively, for each of the remaining candidate samples, the minimum distance to all the already selected samples is computed, so that the one showing the maximum value of this minimum distance is in turn selected to be included in the training set. The whole procedure is then repeated until the desired number of training samples is selected (Westad & Marini, 2015).