Food Chemistry 159 (2014) 458–462

Contents lists available at ScienceDirect

Food Chemistry

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

Analytical Methods

Predicting soluble solid content in intact jaboticaba [*Myrciaria jaboticaba* (Vell.) O. Berg] fruit using near-infrared spectroscopy and chemometrics

Nathália Cristina Torres Mariani [a], Rosangela Câmara da Costa [b], Kássio Michell Gomes de Lima [b,*], Viviani Nardini [a], Luís Carlos Cunha Júnior [a], Gustavo Henrique de Almeida Teixeira [c]

[a] Universidade de São Paulo (USP), Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Departamento de Análises Clínicas, Toxicológicas e Bromatológicas, Av. do Café, s/n, Campus Universitário da USP, Ribeirão Preto, CEP 14040-903 São Paulo, Brazil
[b] Universidade Federal do Rio Grande do Norte, Programa de Pós-Graduação em Química, Grupo de Pesquisa em Química Biológica e Quimiometria, Natal, CEP 59072-970 RN, Brazil
[c] Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Departamento de Produção Vegetal, Via de Acesso Prof. Paulo Donato Castellane, s/n, Jaboticabal, CEP 14884-900 São Paulo, Brazil

ARTICLE INFO

Article history:
Received 21 January 2013
Received in revised form 28 November 2013
Accepted 12 March 2014
Available online 20 March 2014

Keywords:
NIR spectroscopy
PLS
BP-ANN
LS-SVM
Variables selection

ABSTRACT

The aim of this study was to evaluate the potential of near-infrared reflectance spectroscopy (NIR) as a rapid and non-destructive method to determine soluble solid content (SSC) in intact jaboticaba [*Myrciaria jaboticaba* (Vell.) O. Berg] fruit. Multivariate calibration techniques were compared with pre-processed data and variable selection algorithms, such as partial least squares (PLS), interval partial least squares (iPLS), a genetic algorithm (GA), a successive projections algorithm (SPA) and nonlinear techniques (BP-ANN, back propagation of artificial neural networks; LS-SVM, least squares support vector machine) were applied to building the calibration models. The PLS model produced prediction accuracy (R^2 = 0.71, RMSEP = 1.33 °Brix, and RPD = 1.65) while the BP-ANN model (R^2 = 0.68, RMSEM = 1.20 °Brix, and RPD = 1.83) and LS-SVM models achieved lower performance metrics (R^2 = 0.44, RMSEP = 1.89 °Brix, and RPD = 1.16). This study was the first attempt to use NIR spectroscopy as a non-destructive method to determine SSC jaboticaba fruit.

1. Introduction

The jaboticaba has origins in Brazil's south centre. Among the currently known species, the *Myrciaria cauliflora* (MC) and *Myrciaria jaboticaba* (Vell.) Berg, produce fruits that are suitable for both fresh consumption and commercial products. The peel of the mature fruit is black, thin and fragile; the flesh is white, sweet and slightly acidic. The *jabuticabeira* is a fruit tree of great interest to farmers in various regions of Brazil and overseas because of the fruit's high productivity, hardiness and potential uses (Andersen & Andersen, 1988; Teixeira, Durigan, & Durigan, 2011; Teixeira, Durigan, Santos, Hojo, & Cunha, 2011).

The soluble solids content (SSC) is one of the major factors that affects the taste of the fruit and is closely related to the consumer's perception of maturity in jaboticaba fruits. However, most instrumental techniques used to measure SSC are destructive, time-consuming and costly.

Near-infrared (NIR) spectroscopy has been shown to be a rapid and non-destructive method for measuring the internal quality of fruit and is ideally suited to the requirements of the agrofood industry for quality control (He, Zhang, Pereira, Gómez, & Wang, 2005). Research has been conducted in the analysis of SSC by NIR spectroscopy in several fruits, such as apples (Nicolai et al., 2008), kiwifruits (Martinsen & Schaare, 1998), pears (Sun, Lin, Xu, & Ying, 2009), mulberry (Huang et al., 2011), oranges (Liu, Sun, & Ouyang, 2010), and mangos (Jha et al., 2012). The non-destructive technique or chemometric approach, however, has not been reportedly used to analyse SSC in jaboticaba fruit.

The main challenge of NIR spectroscopy is choosing target wavelengths from the full IR spectrum that result in maximum accuracy, especially when the spectra display unresolved peaks or fail to identify important features. To overcome these difficulties, various chemometric algorithms have been proposed to select optimal variables for multivariate calibration, such as iPLS (interval partial least squares) (Norgaard et al., 2000), GA (genetic algorithm) (Ferrand et al., 2011) and SPA (successive projections algorithm) (Araújo et al., 2001), which improve multivariate models by exploiting relevant variables.

Another tool used to improve NIR spectroscopic results is outlier detection, which selects samples that stand out from the bulk data, typically generated by instrumental errors, the presence of

* Corresponding author. Tel.: +55 84 3215 3828; fax: +55 83 3211 9224.
E-mail address: kassio@ufrnet.br (K.M.G. de Lima).

http://dx.doi.org/10.1016/j.foodchem.2014.03.066
0308-8146/

another population or human/analytical errors. In this work, the calibration and prediction sets were optimised based on data with extreme leverage, and unmodeled residuals of the overall data and the dependent variables, specifically (Valderrama, Braga, & Poppi, 2007).

When spectral data present non-linearity that can arise from the chemical nature of the target analytical parameter, linear methods can lead to inaccurate predictions. These errors can be improved by using non-linear multivariate techniques. Least squares support vector machine (LS-SVM), a new learning algorithm proposed by Suykens, Van Gestel, De Brabanter, De Moor, and Vandewalle (2002) is a promising alternative to existing linear and nonlinear multivariate statistical procedures. Another alternative is back-propagation artificial neural networks (BP-ANN) (Hertz, Krogh, & Palmer, 1991), which can also be effective when non-linear spectral responses are involved.

The objectives of this study were (1) to establish the quantitative relationships between the NIR spectra and SSC measurements in intact jaboticaba fruit, and (2) to compare the predictive performances of calibration models established by partial least squares (PLS), back-propagation artificial neural network (BP-ANN), and least squares support vector machine (LS-SVM) on the basis of the selected wavelength variables (iPLS, GA, SPA).

2. Materials and methods

2.1. Fruit samples

A total of 100 jaboticaba fruit [*M. jaboticaba* (Vell) Berg cv. Sabará] were collected in 2011 at the beginning of the harvest season (11/10/2011). The fruits were sorted; 50 were classified as large (10 ± 0.5 g) and 50 as small (5 ± 0.2 g). All fruits harvested were at a stage ready for commercial sale and the skins were completely purple.

2.2. Instrumentation

After temperature stabilization (~25 °C), spectra were collected by a Fourier transform (FT)-IR spectrophotometer (Spectrum 100N, PerkinElmer) in the diffuse reflectance mode over the range of 10,000–4000 cm^{-1} (1000–2500 nm). Spectra were randomly collected from the epidermal surface at two different locations on each fruit (64 scans, spectral resolution of 2 cm^{-1}). The mean spectrum was then calculated for each fruit from the measured spectra.

2.3. Reference methods for SSC

After the spectra were collected, each piece of fruit was analysed for their soluble solids content (SSC) using reference method 920.151 of the Official Methods of Analysis (AOAC, 1997). The fruit was cut in half and one portion was placed in a hydrophilic bandage and pressed to extract the juice. The juice was then placed on a digital refractometer, ATAGO (model PR-101α, Tokyo, Japan) to determine the SSC. This equipment is capable of automatic temperature compensation and a measurement accuracy of ±0.1 °Brix. The descriptive statistics for SSC of the jaboticaba fruit are presented in Table 1.

2.4. Sample distribution

The fruit (100) were assigned to one of two groups as follows: a calibration set and a prediction set. A well-known representative sample selection algorithm, Kennard–Stone, was applied to the sample distribution (Kennard & Stone, 1969). In this study, 70 fruit were used to establish the model. The remaining 30 fruit were used for the prediction. To compare the performance of different calibration models, the fruit in the calibration and prediction sets were the same for all calibration models.

Table 1
Average concentration of soluble solids content (%) of jaboticaba 'Sabará' [*Myrciaria jaboticaba* (Vell.) Berg] fruit collected across three harvests in 2011–2012.

Samples set	*N*	Average	S.D*	Maximum	Minimum
	100	18.9	2.22	23.5	12.1
Large	50	20.0**a**	1.62	22.5	14.8
Small	50	17.7**b**	2.16	23.5	12.1

Averages followed by the same letter, in the column for each harvest, are significantly different according to Tukey's test ($P < 0.05$).
* Standard deviation.

2.5. Spectra pre-processing

Spectral pre-processing can improve the performance of a model. In this study, the reflectance spectrum (R) for each fruit was first log ($1/R$) transformed to give an absorbance spectrum. Several spectral pre-processing algorithms, including Savitzky–Golay smoothing (SGS) over 3–11 points, multiplicative scatter correction (MSC), and 1st and 2nd derivatives were investigated over a span of 3–91 points.

2.6. Multivariate analysis

The calibration spectra were subject to a partial least squares regression (PLSR) with leave-one-out cross validation. The optimal number of latent variables was determined by minimising the predicted residual error sum of squares (PRESS). The PLSR calibration models were evaluated using the coefficient of correlation (R^2) and the root-mean-square error from calibration (RMSEC) and cross validation (RMSECV). In addition, the accuracy of each PLS model was evaluated by comparing their residual predictive deviations (RPDs), which is the ratio of the standard deviation for a specific reference population and RMSEP of the prediction set. According to Nicolaï et al. (2007), a RPD between 1.5 and 2.0 indicates that the model can discriminate low from high values in the response variable. PLS models were calculated using MATLAB version 6.5 (Math-Works, Natick, USA) and the PLS-toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA, version 6.01).

Artificial neural network (ANN) models were built from a reduced number of variables based on the scores from a PCA of the original, smoothed and derived spectra plotted over the wavelength range of 1100–2500 nm. The back-propagating network architecture was subject to supervised training on selected samples known as the monitoring set. The ANN models were obtained from the ANN toolbox in MATLAB version 7.

The least-squares support vector machines (LS-SVM) approach is an optimised version of the standard supporting vector machines. For detailed in-depth theoretical background on LS-SVM, readers are referred to the introduction of Suykens et al. (2002). In this study, the LS-SVM was executed using MATLAB version 7 using the LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) to derive the LS-SVM models.

2.7. Selection variables

The predicted results for the calibration models developed by PLS used the spectral regions selected by iPLS, GA, and SPA. These results were also compared to those found by PLS using the whole region. In the iPLS method, the data were divided to non-overlapping sections where each section undergoes independent PLS