Australian Journal of Crop Science AJCS

AJCS 14(06):889-896 (2020) ISSN:1835-2707
doi: 10.21475/ajcs.20.14.06.p2099

Selection of sugarcane clones via multivariate models using near-infrared (NIR) spectroscopy data

Luiz Alexandre Peternelli[1]*, Mateus Teles Vital Gonçalves[1], Jaqueline Gonçalves Fernandes[2], Bruno Portela Brasileiro[3], Reinaldo Francisco Teófilo[4]

[1]Department of Statistics, Universidade Federal de Viçosa (UFV), 36570-900, Viçosa, Brazil
[2]Departament of Exact Sciences, Universidade Federal de Lavras (UFLA), 37200-000, Lavras, Brazil
[3]Department of Plant Sciences, Universidade Federal do Paraná (UFPR), 80035-050, Curitiba, Brazil
[4]Department of Chemistry, Universidade Federal de Viçosa (UFV), 36570-900, Viçosa, Brazil

***Corresponding author: peternelli@ufv.br**

Abstract

Sugarcane production plays a fundamental role in the Brazilian economy, both for sugar production and renewable energy generation. The development of new cultivars to meet the current needs of the sugarcane industry sector requires efficient phenotyping methods, which should incorporate simplification, speed, accuracy, and consistency. In order to contribute to the development of new phenotyping strategies, this work aimed to develop multivariate regression models using Partial Least Squares (PLS) to classify sugarcane clones based on sugarcane biomass quality parameters, namely fiber (FIB) and apparent sucrose (SC) content. A NIR instrument was used to acquire the reflectance spectra of 196 sugarcane bagasse - collected in two different harvest seasons - and fresh stalk samples. The values predicted by these models allowed the construction of a vector using a confusion matrix that informs whether the clone should be selected or not. PLS models selected to predict each trait under study presented high accuracy and precision, besides small values of false-positive rate and good concordance indication by the Kappa statistic test. The results obtained indicate that the use of fresh stalk samples rather than bagasse samples for the prediction of SC and FIB is recommended as it delivered higher predictive power and is of a more straightforward usage. The utilization of NIR combined with multivariate techniques may help breeding programs in the classification of sugarcane clones based on biomass quality parameters.

Keywords: Calibration models; NIR; PLS; fiber content; apparent sucrose content.
Abbreviations: BBH_ wet bagasse collected at the beginning of the harvest, BMH_ wet bagasse collected at the middle of the harvest, C_o_observed concordance, C_e_expected concordance, D1_first derivative, FIB_fiber content, FP_false positive, FT_Fourrier transform, KS_Kennard and Stone algorithm, LV_latent variables, MC_mean centering, MSC_Multiplicative Scatter Correction, NIR_near-infrared, PCR_principal component regression, PLS_partial least squares, PMGCA-UFV_Sugarcane Genetic Breeding Program of the Universidade Federal de Viçosa, RMSE_root mean squared error, RMSECV_root mean squared error of cross-validation, RMSEP_root mean squared error of prediction, R^2_multiple coefficient of determination, SC_ apparent sucrose content, SMH_stalk collected in the middle of the crop, TN_true negative.

Introduction

There has been a steady increase in investments for the development of alternative energy sources to replace fossil fuels (Zhao et al., 2009). Ethanol obtained from sugarcane juice - first-generation ethanol - is often the major biofuel employed for this purpose in Brazil (Lopes et al., 2016). However, studies are being carried out to harness the lignocellulosic components of the plant aiming to produce ethanol from sugarcane bagasse, the second-generation ethanol (Zheng et al., 2009). Additionally, the cogeneration of electricity can also be conducted by burning the bagasse (Silveira et al., 2015).

Sugarcane mills and distilleries invest in genetic breeding because it allows the obtainment of new and more suitable sugarcane varieties, either sugar production or for renewable energy generation (Silveira et al., 2016). In this sense, sugarcane breeders may be interested in determining sugarcane biomass quality parameters, i.e. fiber or sugar content values at the beginning of a breeding program, to compare clones and to classify them in above or below the overall experimental mean. Moreover, they may be interested in selecting a specific top percentage of the sugarcane clones composing an evaluated population. The clones showing values above the defined threshold are conducted for the next assessment phase of the program, while the rest is discarded. However, the methods commonly used by sugarcane genetic breeding programs for phenotypic evaluation are usually costly and time-consuming. Thus, it is fundamental to develop new phenotyping strategies. The emergence of new technologies has allowed the application of Near Infrared Spectroscopy (NIR), combined with multivariate statistical methods, to determine the biochemical composition of a wide range of plant species biomass feedstock, including sugarcane (Liu et al., 2010; Santchurn et al., 2012; Assis et al., 2017). NIR has

an excellent potential application (Montes. et al., 2013; Roque. et al., 2017) due to its ease of use, speed, accuracy and the absence of waste generation (Pasquini, 2003; Valderrama et al., 2007). The analytical determination of sample's chemical composition using NIR spectroscopy is also a non-destructive methodology as it does not require the use of reagents nor sample preparation (Blanco et al., 2002).

The NIR modeling is carried out using multivariate calibration, usually conducted with the application of Partial Least Squares (PLS) regression or Principal Component Regression (PCR; Beeb et al., 1987), followed by some mathematical pre-treatments of the data (Engel et al.,2013). The PLS regression method is the most recommended in the literature for the analysis of NIR-derived datasets (Brereton, 2000).

The present work aimed to build NIR based multivariate regression models using PLS, as a modern phenotyping strategy to reduce costs and enhance selection efficiency for the prediction and classification of sugarcane clones based on fiber and apparent sucrose content. Specific objectives were to compare the use of sugarcane shred bagasse and fresh stalk samples.

Results and Discussion

NIR spectra analysis

The raw NIR spectra obtained from bagasse samples collected at the beginning of the harvest season, bagasse samples collected in the middle of the harvest season and fresh stalk samples collected in the middle of the harvest season are shown, respectively, in Figure 1-A, 1-B and 1-C. By collecting NIR spectra of sugarcane samples in different seasons, we were particularly interested in evaluating whether the accuracy of the developed models would be affected by the physiological changes on the biomass chemical composition during sugarcane physiological maturation.

Initially, each dataset was sorted into two subsets using the Kennard and Stone algorithm (Kennard. et al., 1969), namely, the calibration set with 166 samples and the test or external validation set with 20 samples. The KS algorithm allows the uniform selection of samples. We assessed the existence of outlier samples present in the data sets through the Leverage versus Studentized residual plot and identified ten samples that were excluded from the analysis (Martens et al., 1992). We performed a leave-one-out (Valderrama et al., 2007) cross-validation in the training set for the selection of the best number of latent variables (LV). This selection was performed based on the analysis of the graph of the Root Mean Squared Error of Cross-Validation (RMSECV) versus LV, which allows the identification of the LV number corresponded to the lowest RMSECV value.

Fiber and apparent sucrose content - chemical analysis results

The chemical analysis indicated that fiber content (FIB) values ranged from 8.38% to 19.51% at the beginning of the harvest (early harvest) and increased in the middle of the harvest ranging between 9.58% and 22.53%. Apparent sucrose content (SC) values ranged from 1.78% to 12.20% at the beginning of the harvest (early harvest) and from 3.11% to 16.89% in the middle of the harvest. These results appear to be related to the sugarcane maturation process for an eighteen-month crop. Sugarcane physiological maturation is the process in which the plant increases the sucrose accumulation in its storage tissues as a response to the changes in the environment (Toppa et al., 2010). The sugarcane biomass chemical composition over the harvest season is a function of several factors, including soil moisture, air temperature, crop management and the genotype (Inman-Bamber et al., 2010; Pereira et al., 2017). However, lower air temperature and water shortages are the most significant contributors to the sugarcane maturation process (Cardozo et al., 2013). At the beginning of the harvest season, adverse weather conditions cause the decrease of the vegetative growth rate and lead to changes in the Carbon partitioning dynamics with lower cell wall synthesis and increase of sucrose accumulation in the stalks (Wang et al., 2013; Botha et al., 2000). Tai et al., (1996) and Wagih et al., (2004) evaluated changes in the accumulation of sucrose and fiber content during sugarcane maturation and observed a steady linear pattern for fiber and a quadratic pattern for sucrose. Likewise, Zhao et al., (2009) found similar results in sorghum, with the biomass components cellulose, hemicellulose and lignin increasing with crop cycle length.

The increase of SC from early to middle harvest can be addressed to a more advanced maturation stage of the sugarcane clones, after undergoing a water shortage caused by the dry month's period (Cardozo et al., 2013; Inman-Bamber et al., 2010). The occurrence of low values of SC is linked to the investigated energy-cane population under study, in which many clones naturally present low SC concentrations. Therefore, since they are aimed for burning or the production of second-generation ethanol, the clones have no aptitude for high sucrose accumulation (Ramos et al., 2017; Legendre et al., 1994).

Fiber and apparent sucrose content prediction

Regarding the FIB trait, according to the lowest RMSECV obtained, we selected nine latent variables (LV) for the bagasse samples dataset collected at the early harvest season, while for the bagasse and stalk samples dataset collected in the middle of the harvest season the lowest RMSECV was associated to four and eight LV, respectively.

Table 2 shows the PLS models built applying different pre-treatments. We tested the combination of different pretreatments and the models' performance regarding the RMSCV and the coefficient of multiple determination of cross-validation. Bagasse samples collected at the beginning (BBH) and in the middle (BMH) of the harvest season did not provide well-fitted models (Table 2). On the other hand, a PLS model built using fresh stalk samples collected in the middle of the harvest season (SMH) without any pre-treatment yielded the best results to predict fiber content, with RMSECV and R^2 values of 1.67 and 0.47, respectively (Table 2). After selecting the best model, we evaluated its performance using the samples that were not included when fitting the training model, i.e. the samples which pertained to the validation set (Pasquini, 2003; James et al.). In this case, the coefficient of multiple determination of prediction obtained was 0.32, and the Root Mean Squared Error of Prediction was 2.7, which is considered a small value since it is nearly four times lower than the least fiber content value

Table 1. Confusion matrix between the correct classification based on the observed values and the classification obtained from the values predicted by the models on the test set.

Classification	SR = 50% (top 10 clones) or 25% (top 5 clones)	
	NS (Predicted)	S (Predicted)
NS (real)	True Negative (TN)	False Positive (FP)
S (real)	False Negative (FN)	True Positive (TP)
Accuracy	(TN+TP)/n	
False-positive rate	FP/(TN + FP)	
Precision	TP/(FP + TP)	
Kappa	$(C_o - C_e)/(1 - C_e)$	

SR: Selection Rate according to the real value (top 25% or top 50%); S: Selected clones; NS: Non-selected clones; C_o: Observed concordance; C_e: Expected concordance, and n: the total number of clones.

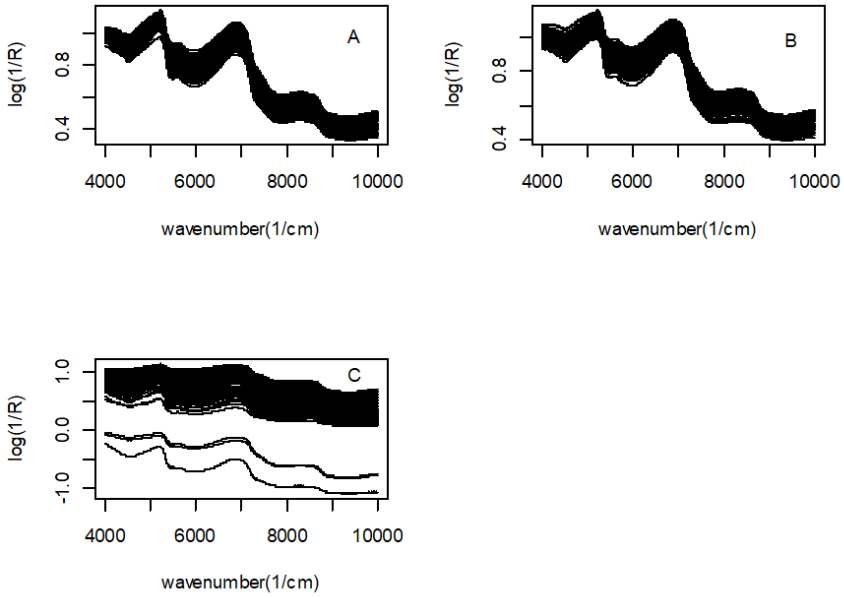

Fig 1. Raw NIR spectra from: (A) bagasse samples collected at the beginning of the harvest season; (B) bagasse samples collected in the middle of the harvest season; and (C) fresh stalk samples collected in the middle of the harvest season.

Table 2. Root Mean Squared Error of Cross-Validation (RMSECV) and coefficient of determination (*R*2) obtained from Partial Least Squares (PLS) analysis using different sugarcane samples and harvest seasons for fiber content.

Pre-treatment	BBH		BMH		SMH	
	RMSECV	R^2	RMSECV	R^2	RMSECV	R^2
None	1.77	0.29	1.95	0.34	1.67	0.47
MC	1.77	0.29	1.97	0.33	2.15	0.28
D1 + MC	2.22	0.06	2.28	0.21	1.78	0.43
MSC + MC	1.85	0.26	2.13	0.27	2.11	0.27
MSC + D1 + MC	2.22	0.06	2.27	0.21	2.06	0.34

BBH: wet bagasse collected at the beginning of the harvest; BMH: wet bagasse collected at the middle of the harvest; SMH: stalk collected in the middle of the harvest; MC: mean centering; D1: first derivative and MSC: Multiplicative Scatter Correction.

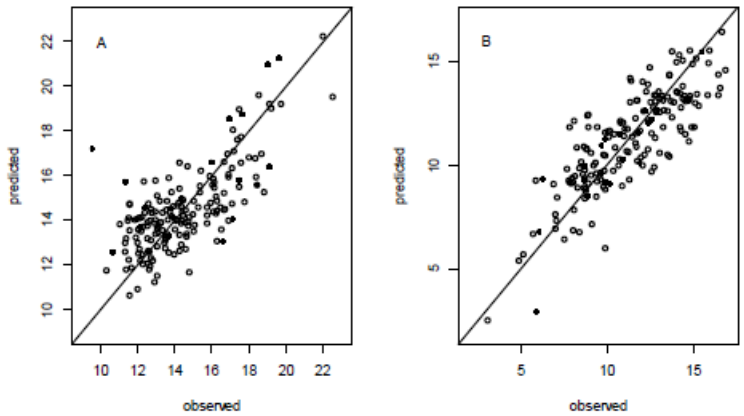

Fig 2. Observed versus predicted: (A) fiber content values obtained from the PLS model built using fresh stalk samples; and (B) apparent sucrose content values obtained from the PLS model built using fresh stalk samples following the application of Multiplicative Scatter Correction, First Derivative and Mean Centering. Empty circles refer to the calibration set; and solid black circles to the test/validation set.

Table 3. Root Mean Squared Error Statistics by Cross-Validation (RMSECV) and coefficient of determination (*R*2) obtained from the Partial Least Squares (PLS) model in different samples and harvest seasons for apparent sucrose content.

Pre-treatment	BBH		BMH		SMH	
	RMSECV	R^2	RMSECV	R^2	RMSECV	R^2
None	1.72	0.54	2.01	0.47	3.28	0.20
MC	1.77	0.52	2.05	0.44	2.94	0.27
D1 + MC	1.79	0.52	2.05	0.19	1.83	0.58
MSC + MC	1.68	0.57	1.89	0.52	1.73	0.61
MSC + D1 + MC	1.81	0.52	3.05	0.05	1.56	0.68

BBH: wet bagasse collected at the beginning of the harvest; BMH: wet bagasse collected at the middle of the harvest; SMH: stalk collected in the middle of the harvest; MC: mean centering; D1: first derivative and MSC: Multiplicative Scatter Correction.

Table 4. Confusion matrix between the correct classification based on the real values of fiber content (FIB) and the classification obtained by the predicted values from the PLS analysis. Models constructed from fresh stalk samples without any pre-treatment.

Classification	SR = 50% (top 10 clones)		SR = 25% (top 5 clones)	
	NS (predicted)	S (predicted)	NS (predicted)	S (predicted)
NS (real)	8	2	13	2
S (real)	2	8	2	3
Accuracy	0.80		0.80	
False positive rate	0.20		0.13	
Precision	0.80		0.60	
Kappa	0.60 (p-value: 0.037)		0.47 (p-value: 0.007)	

SR: Selection rate based on the real value; S: Selected clones, and NS: Non-selected clones.

Table 5. Confusion matrix between the correct classification based on the observed values of the apparent sucrose content (SC) and the classification obtained from the PLS model constructed using fresh stalk samples after the application of the Multiplicative Scatter Correction, First Derivative and Mean Centering.

Classification	SR = 50% (top 10 clones)		SR = 25% (top 5 clones)	
	NS (predicted)	S (predicted)	NS (predicted)	S (predicted)
NS (real)	8	2	14	1
S (real)	2	8	1	4
Accuracy	0.80		0.90	
False-positive rate	0.20		0.06	
Precision	0.80		0.80	
Kappa	0.60 (p-value: 0.037)		0.73 (p-value: 0.001)	

SR: Selection Rate based on the real value; S: Selected clones and NS: Non-selected clones.

(Ferreira, 2105). Figure 2-A shows the graph of the observed versus predicted values of FIB using fresh stalks samples.

Besides FIB, we also attempted to develop regression models to predict SC values using sugarcane bagasse and fresh stalks samples. After analyzing the LV × RMSECV relationship, we selected eight LV for the BBH samples, and five and seven LV, respectively, for BMH and SMH samples. In both scenarios, the prediction of FIB and SC, the number of LV agrees with the maximum number suggested by Pasquini (2003). Again, the model built using SMH samples yielded the best results for the prediction of SC, with the lower value of RMSECV (1.56) and the highest value of R^2 (0.68; Table 3). The results observed herein suggest that, regarding the SC trait, it is also recommended to use NIR spectra collected from fresh stalk samples after applying the pre-treatments MSC, 1st Derivative and Mean Centering. For the prediction of SC, the removal of the outlier samples identified from the Leverage vs. Standardized Student Residues plot improved the performance of the selected model. The R^2 increased from 0.68 to 0.71, and the RMSECV decreased from 1.56 to 1.48 (Martens et al., 1992). Again, after selecting the best-fitted model, we tested its predictive ability in the samples of the validation set The values of R^2 and RMSEP obtained were 0.64 and 3.07, respectively (Table 3). Figure 2-B shows the graph of the observed versus predicted values of SC using fresh stalks samples.

Alongside with inherent composition variations present in the samples, variable spectral path length and radiation scattering due to different particle size may negatively influence the analysis of NIR diffuse reflectance spectra (Ely et al., 2008; Barnes et al., 1989). Therefore, these physical interferes may conceal the chemical information the researcher was initially investigating and ultimately affect the models' prediction power (Engel et al., 2013). In this sense, the shred bagasse samples characterized by rough and different particle sizes may have been the cause of the underperformance compared to the models built using fresh stalk samples. However, Sabatier et al., (2011) stated that the mathematical pre-treatments commonly applied in NIR spectra were enough to correct spectra variations resulted from sugarcane samples with coarse particle size. Nevertheless, the utilization of fresh stalk over bagasse samples is advantageous as it requires less sample preparation. Thus it brings greater ease to the analysis.

Assis et al., (2017) successfully predicted lignin content of sugarcane fresh stalk samples using NIR spectroscopy, obtaining a higher accuracy than we found herein. Moreover, results obtained by Valderrama et al., (2007) using a NIR protocol to predict SC in sugarcane juice samples also demonstrate the possibility of developing NIR based models with higher accuracies. However, in this study, we were not only concerned with the accuracy of predictions;

we also propose the classification of sugarcane clones using the NIR based developed models.

Fiber and apparent sucrose content classification

In this section, we computed a confusion matrix for the test set by assuming the selection of 10 (50%) and 5 (25%) of the top-ranked sugarcane clones based on FIB and SC values of the total 20 samples of the validation set.

Table 4 shows the confusion matrix for the classification of clones based on FIB, after considering a threshold for the selection of the 10 and 5 top-ranked clones. From Table 4 we see that the PLS model returned predicted values capable of classifying clones in selected or non-selected with 80% accuracy, regardless of the selection rate considered. The Kappa test showed significance on both scenarios (SR = 50% and 25%), indicating an excellent performance of the model for classifying clones in selected or non-selected categories. The results from Table 4 also indicate that the classification model presented small values of false-positive rate, with values of 0.13 and 0.20 for the selection rates 50% and 25%, respectively.

Table 5 shows the confusion matrix for the classification of clones based on SC, after considering a threshold for the selection of the 10 and 5 top-ranked clones. Despite presenting a moderate value of R^2 (0.65), this model might be a useful tool for ranking sugarcane clones. It is highly accurate (accuracy of 0.8 and 0.9, depending on the number of clones selected) and presents low values for false-positive rate (0.2) and it indicates, according to the Kappa test ($p < 0.05$), the existence of agreement in the classification of the clones based on the training model (Table 5).

It is essential to highlight the need for interpretation the false positive rate results in this study (James et al., 2013). It relies on the fact that even if the breeder takes a below-average clone and conduct it further in a breeding program because the fitted model erroneously indicated it, it is still possible to discard this clone in the next year when this clone will undergo a new evaluation trial. Therefore, at this moment, we are not only concerned about the overall error rate obtained on the confusion matrix, but also on the number of false positives eventually indicated.

The PLS models adjusted for the NIR data presented a high potential for classifying sugarcane clones based on biomass quality parameters. Similar results were also found in other studies that used similar methodologies (Montes et al., Roque et al.,2017). The results observed are encouraging and suggest that the models developed in this study can be used as a screening tool to aid breeder's decision in the selection of sugarcane clones.

Materials and Methods

Plant material and phenotypic evaluation

The Sugarcane Genetic Breeding Program of the Universidade Federal de Viçosa (PMGCA-UFV) consists of five phases: first, second, and third testing phases; multiplication phase; and experimental phase (Barbosa et al., 2012). After performing clonal selection in the best families evaluated in the first testing phase, 196 clones were submitted to the second testing phase. The selected clones are contrasting in fiber content (FIB) (Rocha et al., 2012; Souza et al.,2013) and apparent sucrose content (SC) (Fernandes, 2011) traits. These clones were originated from crosses involving parents with high sucrose content and with high fiber content, to obtain energy cane cultivars that combine both traits.

The second testing phase was installed in July 2014 in augmented block design, with two cultivars as checks (RB867515 and C90-176) with 18 blocks in total. Each of the 196 plots consisted of two 5-meters-long furrows, spaced 1.4 m. We installed an experiment at the PMGCA-UFV Research Center, located in the municipality of Oratórios, MG, the latitude of 20°25'; longitude of 42°48'; altitude of 494 m and LVE soil.

We obtained FIB and SC measures through the technical analysis of 500 g samples of wet bagasse obtained from the milling of 10 sugarcane stalks per plot in two seasons: early harvest at ten months (May 2015) after planting; and middle harvest at 13 months (August 2015) after planting.

Collection of the NIR spectra data

We collected NIR data from bagasse samples collected on the early and middle harvest, and from fresh stalk samples collected on the middle harvest, as explained below.

Around 200 g of shred bagasse, obtained from the milling of 10 stalks of each of the 196 sugarcane samples were frozen at -20ºC to avoid deterioration until one could perform the read the NIR samples in the laboratory. After 30 days of storage, each sample was thawed and approximately 3 g of wet bagasse was placed into a recipient. After, we performed NIR readings using an Agilent 660 Fourier transform (FT) spectrometer. We did three readings on each sample, at each reading we gently moved the recipient containing the bagasse so that a new reading could be carried out in another part of the sample. After three days, we completed all the 588 readings. We used the same Agilent 660 Fourier transform (FT) spectrometer to acquire all fresh stalk samples spectra. In this case, we froze at -20°C three internodes of the middle third of three stalks (nine internodes) per plot. After 30 days, we thawed the internodes of the 196 stalk samples. Each internode was cut lengthwise and used for a single reading in the NIR instrument.

Statistical analysis

We organized the spectral data in an **X** matrix, in which the rows correspond to the sugarcane samples, and columns corresponded to the covariates, i.e., the NIR wavenumbers. The response vector **y** contained the values of each evaluated trait, namely FIB, and SC.

Before the analysis, we submitted the data to three different types of pre-treatments: first derivative, multiplicative scatter correction and mean centering (Engel et al., 2013). The first derivative has the purpose of correcting the changes in the baseline due to the instrument or sampling systematic variations. We employed the Savitzky-Golay smoothing method using a 15 sized window and a second-degree polynomial. The Multiplicative Scatter Correction (MSC) aims to correct the effect of light scattering due to the lack of particle size and distribution homogeneity in the samples or the variations resulting from differences in the optical path length of the samples (Rinnan et al., 2009). The mean centering aims to give more relevance to the distance of the points to the mean value and eliminate from the data the intensity value of each variable (Rinnan et al., 2009).

The Partial Least Squares (PLS) regression method was employed. This method has the advantage of considering the information of the response variable vector *y* for the

construction of the model (Wold et al., 2001). Moreover, it can cope with highly correlated data, thus improving the representation of the information contained in the NIR spectrum (Brereton et al., 2018)

The PLS fitting procedure can be developed based on the bidiagonal algorithm (Barlow.et al., 2005), corresponding to the decomposition of an X matrix into three other matrices, according to the equation below:

$$\mathbf{X} = \mathbf{URV^t}$$

where **X** is the data matrix already incorporated with the vector **y**; **R** is a diagonal matrix; **U** and **V** are matrices organized so that their first columns are composed of the information present in **X** in decreasing order, called latent variables. Thus, much of the information can be reconstructed from **X** with only part of **U** and **V**.

The coefficient of multiple determination (R^2) and the Root Mean Squared Error (RMSE) were employed as the evaluation criteria, which can be computed as follows:

$$R^2 = \frac{\left[\sum_{j=1}^{n}(\hat{y}_j - \bar{\hat{y}})(y_i - \bar{y})\right]^2}{\sum_{j=1}^{n}(\hat{y}_j - \bar{\hat{y}})^2 \sum_{j=1}^{n}(y_i - \bar{y})} \quad RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

where y_i corresponds to the real value of sample *i* (*i* = 1 to *n*); $\hat{y}_i$ corresponds to the predicted values for sample *i*, and *n* is the total number of samples considered in the calculation. The RMSE can be calculated using $\hat{y}_i$ as the values obtained in cross-validation (RMSECV) or using $\hat{y}_i$ as the prediction values (RMSEP). We considered a model adequate when the values of RMSECV and RMSEP are sufficiently small, that is, smaller than the minimum value of the dependent variable, combined with the highest R^2. In addition to the performance of regressions, RMSECV is employed to select the optimal number of latent variables for the models (Martens et al.,1992). We used the PLS-Toolbox 4.0 algorithm package for the analysis, in the Matlab software, version 6.0 (The Mathworks, Natick, USA) to perform all statistical analysis.

Technical Assessment and comparisons

Due to the relevance of knowing the best clones for genetic breeding programs, we classified the clones by ordering the values of fiber content (FIB), and apparent sucrose content (SC) measured. We investigated two scenarios regarding the selection rates. In the first scenario we fixed a selection rate of 50% and of 25% in the second. A binary data vector was created for each trait (FIB and SC), assuming a value equal to 1 if the clone was selected and a value of 0, otherwise. The selection is performed for clones with high FIB and also high SC. The same classification was carried out based on the results obtained by the fitted model.

The binary values obtained from the training set in which the model was fitted and validation set were the basis for the computation of the confusion matrix, also known as contingency matrix (James et al., 2013; Table 1). The confusion matrix evaluation criteria considered the measurements of accuracy, false-positive rate, precision and Kappa concordance test (Fleiss.et al., 1981; Castellan, 1988) obtained from this matrix.

Conclusion

The prediction of fiber and apparent sucrose content can be performed from stalk samples instead of wet bagasse samples due to its higher predictive power and ease of applicability when considering NIR readings. Also, the employment of a protocol to screen sugarcane clones using NIR spectroscopy may save a significant amount of resources as ordinary phenotyping strategies currently adopted represent a high-cost element in the PMGCA. In the population under study, we achieved the best results for the data collected in the middle of the harvest season. For fiber content, no pre-treatment was necessary to obtain the best model, whereas, for apparent sucrose content, it was necessary to apply the pre-treatments Multiplicative Scatter Correction, First Derivative and Mean Centering. The models used to select the top clones regarding fiber and sucrose content showed high accuracy, high precision, and low values of false-positive rates. Therefore, the results obtained in this study suggest that the use of NIR combined with multivariate techniques may help breeding programs on classifying and selecting sugarcane clones efficiently.

Acknowledgments

This study was partly financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also thank the Foundation for Research of the State of Minas Gerais (FAPEMIG) for the financial support of research projects and the National Council for Scientific and Technological Development (CNPq) for the research scholarships. Finally, we thank RIDESA, the Inter-University Network for the Development of the Sugarcane Industry in Brazil, for providing support on the field experiments.

References

Assis C, Ramos RS, Silva LA, Kist V, Barbosa MHP, Teófilo RF (2017) Prediction of lignin content in different parts of sugarcane using near-infrared spectroscopy (NIR), ordered predictors selection (OPS), and partial least squares (PLS). Applied Spectroscopy. 0(0):1-12.

Barbosa MHP, Silveira LCI (2012) Breeding and Cultivar Recommendations. In: Santos F, Borém A, Caldas C (eds) Sugarcane: Bioenergy, Sugar, and Ethanol—Technology and Prospects. Suprema, Vicosa, MG, 568 p.

Barlow JL, Bosner N, Drmac Z (2005) A new stable bidiagonal reduction algorithm.

Linear Algebra and its Applications. 397: 35–84.

Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied spectroscopy. 43(5): 772-777.

Beebe KR, Kowalski BR (1987) An introduction to multivariate calibration and analysis.

Analytical Chemistry. 59(17): 1007A–1017A.

Botha FC, Black KG (2000) Sucrose phosphate synthase and sucrose synthase activity during maturation of intermodal tissue in sugarcane. Australian Journal of Plant Physiology. 27: 81-85.

Blanco M, Villarroya I (2002) NIR spectroscopy: a rapid-response analytical tool. TrAC Trends in Analytical Chemistry. 21(4): 240–250.

Brereton RG (2000) Introduction to multivariate calibration in analytical chemistry electronic. Analyst. 125(11): 2125–2154.

Brereton, RG, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O, Roger JM, Walczak B, Tauler R (2018) Chemometrics in analytical chemistry – part II: modeling,

validation, and applications. Analytical and Bioanalytical Chemistry. 410:6691-6704.

Cardozo NP, Sentelhas PC (2013) Climatic effects on sugarcane ripening under the influence of cultivars and crop age. Scientia Agricola. 70(6): 449-456.

De Souza MS, Amanda P (2017) Advances of Basic Science for Second Generation Bioethanol from Sugarcane. 1rst edn. Springer.

Ely DR, Thommes M, Carvajal MT (2008) Analysis of the effects of particle size and densification on NIR spectra. Colloids and Surfaces A: Physicochemical and Engineering Aspects. 331:63-67.

Engel J, Gerretzen J, Szymanska E, Jansen JJ, Downey G, Blanchet L, Buydens LMC (2013) Breaking with trends in pre-processing? Trends in Analytical Chemistry. 50:96-106.

Fernandes AC (2011) Cálculos na agroindústria da cana-de-açúcar. STAB: Piracicaba, SP, Brasil. 3rd edn. 416p.

Ferreira MMC (2015) Quimiometria – Conceitos, métodos e aplicações. Unicamp Campinas, Brazil 1st edn. 493p.

Fleiss JL, Levin B, Paik MC (1981) The analysis of data from matched samples. Statistical Methods for Rates and Proportions, 3rd edn. Wiley Online Library.

Inman-Bamber NG, Bonnett GD, Spillman MF, Hewitt MH, Glassop D (2010) Sucrose accumulation in sugarcane is influenced by temperature and genotype through the carbon source-sink balance. Crop & Pasture Science. 61:111-121.

Jackson JE (2005) A user's guide to principal components. 2nd ed. John Wiley & Sons, United States of America.

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R. Springer Texts in Statistics, New York, 1rst edn. 440p.

Kennard RW, Stone LA (1969) Computer-aided design of experiments. Technometrics.11(1): 137–148.

Lopes ML, Paulillo SCL, Godoy RAC, Lorenzi MS, Giometti FHC, Bernardino CD, Neto HBA, Amorim HV. Ethanol production in Brazil: a bridge between Science and industry. Brazilian Journal of Microbiology. 47:64-76.

Liu L, Ye P, Womac AR, Sokhansanj S (2010) Variability of biomass chemical composition and rapid analysis using FT_NIR techniques. Carbohydrate Polymers. 81: 820-829.

Martens H, Naes T (1992) Multivariate calibration. 1st edn. Wiley, New York.

Montes JM, Technow F, Bohlinger B, Becker K (2013) Grain quality determination by means of near-infrared spectroscopy in jatropha curcas l. Industrial crops and products. 43: 301–305.

Pasquini C (2003) Near-infrared spectroscopy: fundamentals, practical aspects, and analytical applications. Journal of the Brazilian Chemical Society. 14(2): 198–219.

Pereira LFM, Ferreira VM, Oliveira NG, Sarmento LVS, Endres L, Teodoro I (2017Sugar levels of four sugarcane genotypes in different stem portions during the maturation phase. Annals of the Brazilian Academy of Sciences. 89(2): 1231-1242.

Ramos RS, Brasileiro BP, Kist V, Assis, Gasparini K, Silva LA, Teófilo RF, Peternelli LA, Barbosa MHP (2017) Selection of energy cane clones. Crop Breeding and Applied Biotechnology. 17:327-333.

Rinnan Å, Berg FVD, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry. 28(10): 1201–1222.

Rocha GJM, Martín C, Silva VFN, Gómez EO, Gonçalves AR (2012) Mass balance of pilot-scale pretreatment of sugarcane bagasse by steam explosion followed by alkaline delignification. Bioresource technology111: 447–452.

Roque JV, Dias LAS, Teófilo RF (2017) Multivariate calibration to determine phorbol esters in seeds of *Jatropha curcas L*. using near-infrared and ultraviolet spectroscopies. Journal of the Brazilian Chemical Society. 28(8): 1506–1516.

Sabatier D, Dardenne P, Thuriès J (2011) Near-infrared reflectance calibration optimization to predict lignocellulosic compounds in sugarcane samples with coarse particle size. Journal of near-infrared spectroscopy. 19:199-209.

Santchurn D, Randoyal K, Houssen BMG, Labuschagne M (2012) From the sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica. 185(3): 543–558.

Silveira LCI, Brasileiro BP, Kist V, Weber H, Daros E, Peternelli LA, Barbosa MHP (2016) Selection in energy cane families. Crop Breeding and Applied Biotechnology. 16(4): 298–306.

Silveira LCI, Brasileiro BP, Kist V, Weber H, Daros E, Peternelli LA, Barbosa MHP (2015) Selection strategy in families of energy cane based on biomass production and quality traits. Euphytica, 204(2): 443–455.

Souza AP, Leite DCC, Pattathil S, Hahn MG, Buckeridge MS (2013) Composition and structure of sugarcane cell wall polysaccharides: implications for second-generation bioethanol production. BioEnergy Research, Springer, 6(2): 564–579.

Tai PY, Powell J, Perdomo R, Eiland B (1996) Changes in sucrose and fiber contents during sugarcane maturation. Sugar Cane, 6(1): 19-23.

Toppa EVB, Jadoski CJ, Hulshof T, Ono EO, Rodrigues, JD (2010) Physiology aspects of sugarcane production. Applied Research & Agrotechnology. 3(3): 223–230.

Valderrama P, Braga JW, Poppi, RJ (2007) Validation of multivariate calibration models in the determination of sugar cane quality parameters by near-infrared spectroscopy. Journal of the Brazilian Chemical Society. 18(2): 259–266.v

Wagih ME, Ala A, Musa Y (2004) Evaluation of sugarcane varieties for maturity earliness and selection for efficient sugar accumulation. Sugar Tech. 6(4): 297-304.

Wang J, Nayak S, Koch K, Ming R (2013) Carbon partitioning in sugarcane (*Saccharum* species). Frontiers in Plant Science. 4(201): 1-6

Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 58:109-130.

Zhao YL, Dolat A, Steinberger Y, Wang X, Osman A, Xie GH (2009) Biomass yield and changes in the chemical composition of sweet sorghum cultivars grown for biofuel. Field and Crops Research. 111:55-64.

Zheng Y, Pan Z, Zhang R, Wang D (2009) Enzymatic saccharification of dilute acid pretreated saline crops for fermentable sugar production. Applied Energy. 86(11): 2459–2465.