Phenotypic classification of sugarcane from near infrared spectra obtained directly from stalk using ordered predictors selection and partial least squares-discriminant analysis

L.A. Peternelli,[a]* M.H.P. Barbosa,[b] J.V. Roque[c] and R.F. Teofilo[c]

[a]Department of Statistics, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil. E-mail: peternelli@ufv.br
[b]Department of Agronomy, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil
[c]Department of Chemistry, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil

A new method was developed for the early selection of sugarcane genotypes using near infrared spectroscopy combined with partial least squares discriminant analysis (PLS-DA) and a variable selection method named ordered predictors selection (OPS). The use of the OPS method improved the predictive capacity of PLS-DA models to classify the sugarcane samples correctly according to fiber content (FC) and pol percent (PP).

Introduction

The search for the development of renewable fuels is rapidly increasing due to greater concerns about the environmental problems caused by fossil fuels. Biofuels, such as biodiesel and bioethanol, are products that combine energy security and sustainability.[1] Sugarcane (*Saccharum spp.*) is an important alternative energy source and can be considered one of the most important crops in the current Brazilian national farming business scenario.[2]

The production of sugarcane plays a fundamental role in the economy of several countries. However, the average productivity of the cultivars is far below what can be achieved with the genetic potential of the crop, which justifies the need for optimizing breeding. Because the release of new cultivars has occurred 11–13 years after a new breeding cycle starts, phenotyping becomes extremely costly due to the large number of experiments required for the evaluation of clones. Therefore, there is a growing search for alternative methods for phenotyping.

Fiber content (FC) and pol percent (PP) are critical parameters that are often used to select sugarcane genotypes in breeding programs. High amounts of these parameters are desirable in breeding due to their potential use in biofuels production. A higher FC corresponds to a higher biomass volume in the process of producing second generation ethanol and electricity. A higher PP corresponds to a higher sucrose yield for sugar and ethanol production. However, conventional methods to quantify FC and PP are time consuming and destructive. In turn, spectroscopic techniques are non-destructive and there is no reagent consumption or waste generation.[3] Near infrared (NIR) spectroscopy combined with multivariate regression methods provide fast, reliable and accurate methods for classifying complex samples.

This work aims to present a new method for the early selection of sugarcane genotypes by determination of FC and PP using NIR combined with partial least squares discriminant analysis (PLS-DA) and a method of variable selection, the ordered predictors selection (OPS).[4,5] NIR spectra were obtained directly on the stalk, which is the main innovation of this work.

Correspondence
L.A. Peternelli (peternelli@ufv.br)
doi: 10.1255/nir2017.157
Citation: L.A. Peternelli, M.H.P. Barbosa, J.V. Roque and R.F. Teofilo, "Phenotypic classification of sugarcane from near infrared spectra obtained directly from stalk using ordered predictors selection and partial least squares-discriminant analysis", in *Proc. 18th Int. Conf. Near Infrared Spectrosc.*, Ed by S.B. Engelsen, K.M. Sørensen and F. van den Berg. IM Publications Open, Chichester, pp. 157–161 (2019). https://doi.org/10.1255/nir2017.157

ISBN: 978-1-906715-27-4

Material and methods

Sugarcane samples were supplied by the germplasm bank of the Sugarcane Genetic Breeding Program (PMGCA) from Universidade Federal de Viçosa, Viçosa, Minas Gerais (MG), Brazil. A total of 168 samples of sugarcane stalks were analyzed in this work.

Reflectance NIR spectra were collected using a Fourier transform near infrared (FT-NIR) spectrometer (Thermo Scientific Antaris II) controlled with TQ Analysis software. Stalk scans were the average result of 32 scans measured with $4\,\mathrm{cm}^{-1}$ resolution over the wavenumber range 10,000–4000 cm^{-1}. These spectra were obtained in reflectance mode as log (1/R), where R is the collected reflectance. Sample spectra were collected in duplicate and the average spectrum was used in data analysis.

Chemical analysis of FC and PP were carried out according to reference analysis in the literature. The FC contents (%) and PP were measured by gravimetry[6] and polarimetry,[7] respectively.

All calculations were performed in Matlab (Matlab R2016a, 9.0, The MathWorks Inc., Natick, USA) and PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, USA). The OPS® Toolbox is available on the internet at http://www.deq.ufv.br/chemometrics. The original data set was then split in training and test sets using the Kennard–Stone[8] algorithm. The training and test sets were, respectively, 139 and 20 samples for FC, and 148 and 20 samples for PP. Before fitting the classification models, the NIR spectra were pretreated using second and first derivative for FC and PP models, respectively, followed by mean centering.

Classification models were built using the PLS-DA method and a variable selection was performed by OPS. PLS-DA is an adaptation of PLS regression to the problem of supervised clustering.[9] The OPS method is based on obtaining an informative vector that contains information about the location of the best response variables for prediction. Briefly, the original independent variables ($\mathbf{X}$ matrix columns) are differentiated according to the corresponding absolute values of the informative vector elements. The informative vector can be obtained directly from calculations performed with response ($\mathbf{X}$ matrix columns) and dependent variables ($\mathbf{y}$). This vector can be obtained using the regression coefficients (REG), the correlation between each column of matrix $\mathbf{X}$ with $\mathbf{y}$ (COR), residual information of reconstructed matrix with *h* latent variables (SQR), variable importance on projection (VIP), net analyte signal (NAS) and covariance procedures (COV). Details about these vector calculations can be found elsewhere.[4,5] The differentiated variables are sorted in descending order. First, a subset of 20 variables (window) is selected to build and evaluate the first model. Then, this matrix is expanded by the addition of five variables (increment) and a new model is built and evaluated. Quality parameters of the models are obtained for every evaluation. Finally, the evaluated variable sets (initial window and its extensions) are compared using the quality parameters calculated during validations.[4,5]

In this study, stalk samples were classified according to a range of FC and PP determined by reference analysis. Samples were defined in two classes for FC and PP models. For FC, the first class (class 1) has low fiber content with values $7.0 < \mathrm{FCa} < 13.400$; and the second class (class 2) has high fiber content with values $13.401 < \mathrm{FCb} < 20$. For PP classes, the first one (class 1) was defined with low pol percent with values $1.5 < \mathrm{PPa} < 7.50$; and the second one (class 2) with high pol percent with values $7.51 < \mathrm{PPb} < 13.0$.

The classification performances were defined using Receiver Operating Characteristics (ROC) curves, where the sensitivity and the specificity of the model can be accessed. The sensitivity is the percentage of samples of each class accepted by the class model, and specificity is the percentage of samples of the other classes correctly rejected by the class model.[9] In practice, a threshold is determined, above which a sample is in the class and below which a sample is not in the class. A perfect diagnostic test will have a sensitivity and specificity of 1.

Results and discussion

PLS-DA results for Full models, i.e., models using all variables, and for OPS models are showed in Table 1. PLS-DA-OPS models showed a significant improvement in relation to the model that used all the variables. Besides that, for FC classes, OPS models showed excellent predictability ($\mathit{Specificity}_{(pred)} = 1$). The correlation coefficients of calibration (Rcal), cross-validation (Rcv) and prediction (Rpred) are considerably higher for PLS-DA-OPS than for PLS-DA-Full models. In Figure 1, PLS-DA-Full models (A and B) did not provide a proper separation of samples for FC and PP classes, while PLS-DA-OPS models (C and D) could predict the samples correctly. Thus, the model that used selected variables was considered more efficient and robust.

Table 1. Parameters of PLS-DA models using all variables (Full) and variables selected by OPS.

	FC				PP			
	FULL		OPS		FULL		OPS	
Variables	3113		71		3113		147	
nLV	4		4		6		6	
	FCa	FCb	FCa	FCb	PPa	PPb	PPa	PPb
Specificity (Cal)	1.00	0.94	1.00	0.98	0.93	0.89	1.00	0.98
Specificity (CV)	0.55	0.59	0.97	0.97	0.80	0.70	0.93	0.93
Specificity (Pred)	0.50	0.80	1.00	1.00	0.60	0.90	0.90	0.80
Class Err. (Cal)	0.0270		0.0067		0.0870		0.0064	
Class Err. (CV)	0.4258		0.0289		0.2474		0.0670	
Class Err. (Pred)	0.3333		0.0000		0.2500		0.1500	
R (Cal)	0.80		0.89		0.56		0.80	
R (CV)	0.03		0.77		0.32		0.70	
R (Pred)	0.21		0.87		0.31		0.60	

nLV: number of latent variables; Cal: calibration; CV: cross-validation; Pred: prediction; R: correlation coefficient

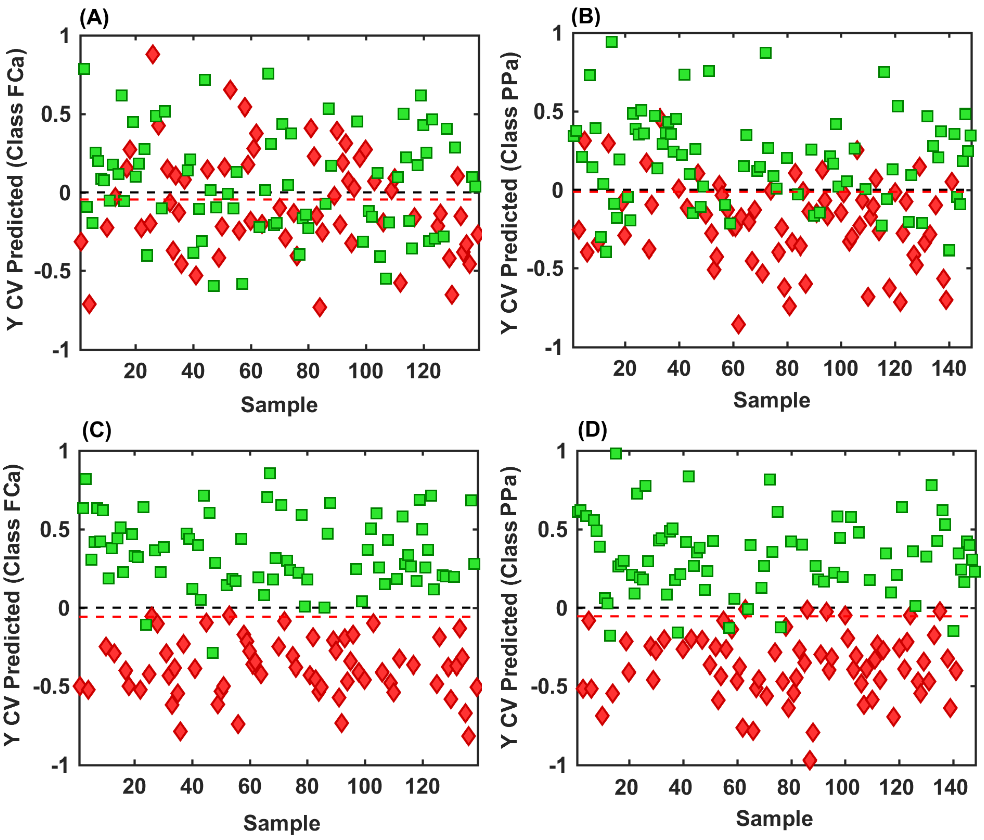

Figure 1. PLS-DA model predictions for class FCa in PLS-DA-Full (A) and PLS-DA-OPS (C) models. PLS-DA model predictions for class PPa in PLS-DA-Full (B) and PLS-DA-OPS (D) models. Green squares are the classes FCa and PPa and red diamonds are the classes FCb and PPb. The horizontal dashed red line indicates the threshold selected by the PLSDA algorithm.

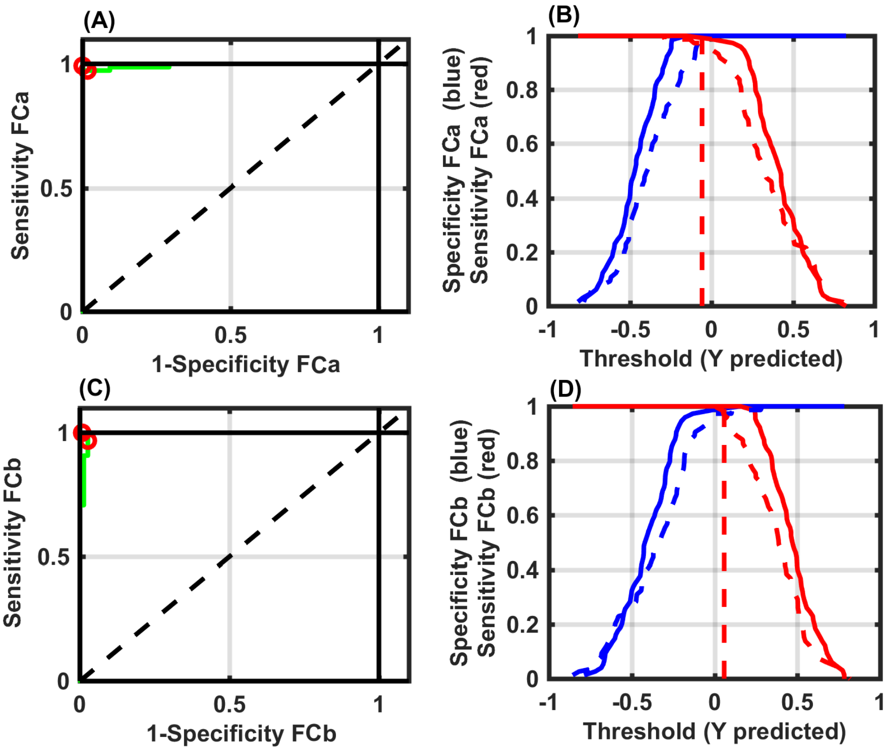

Figure 2. Sensitivity and specificity of FC classes of PLS-DA-OPS models. (A–C) specificity vs sensitivity, (B–D) specificity and sensitivity vs threshold. The vertical dashed red line indicates the threshold selected by the PLSDA algorithm.

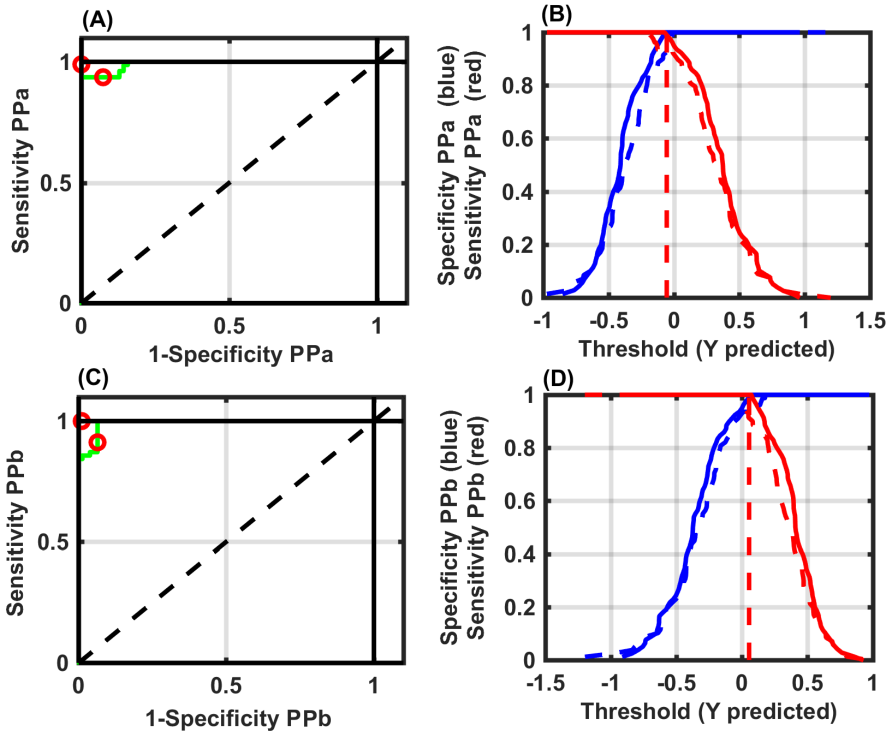

Figure 3. Sensitivity and specificity of PP classes of PLS-DA-OPS models. (A–C) specificity vs sensitivity, (B–D) specificity and sensitivity vs threshold.

The ROC curves for each class are shown in Figure 2 (A and C) and Figure 3 (A and C) for FC and PP, respectively. The diagonal divides the ROC space into two regions, where points above the diagonal represent good classification results and points below the line represent poor results. For both classes of FC and PP, the results were good for classification purposes. Ideally, a curve which reaches the upper-left corner implies that at some threshold, the specificity could be perfect without loss of sensitivity.[9]

In Figure 2 (B and D) and Figure 3 (B and D) the sensitivity and specificity are shown as the threshold value is varied for FC and PP, respectively. Ideally, these lines cross while still at a value of 1. Crossing below a value of 1 indicates that, as the threshold is increased, sensitivity begins to decrease before the model is entirely specific.[9] The vertical dashed red line indicates the threshold selected by the PLS-DA algorithm. So, in our PLS-DA-OPS models there is a loss of sensitivity to achieve more specificity.

Therefore, the results show the feasibility of NIR, PLS-DA and OPS to classify stalk samples of the early selection of sugarcane genotypes.

Conclusions

This study has proven that NIR spectroscopy combined with PLS-DA is a very powerful tool for classifying sugarcane stalk samples according to fiber content and pol percent. PLS-DA-OPS models showed better results than models obtained using all variables. In addition, NIR spectroscopy is an advantageous technique because it is non-destructive, rapid and not expensive. Thus, the methods developed in this work are simple, require just a few steps and can be perfectly applied as alternative methods to select genotypes in breeding programs of sugarcane.

Acknowledgments

The authors are thankful to RIDESA for providing the data and fieldwork support, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq - processes 310303/2015-0 and 310503/2015-9, and FAPEMIG for the financial support for research projects.

References

1. C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa and R.F. Teofilo, "Prediction of lignin content in different parts of sugarcane using near-infrared spectroscopy (NIR), ordered predictors selection (OPS) and partial least squares (PLS)", *Appl. Spectrosc.* **71,** 1–12 (2017). https://doi.org/10.1177/0003702817704147
2. I.P. Caliari, M.H.P. Barbosa, S.O. Ferreira and R.F. Teófilo, "Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods", *Carbohydr. Polym.* **158,** 20–28 (2017). https://doi.org/10.1016/j.carbpol.2016.12.005
3. C. Pasquini, "Near-infrared spectroscopy: fundamentals, practical aspects, and analytical applications", *J. Braz. Chem. Soc.* **14,** 198–219 (2003). https://doi.org/10.1590/S0103-50532003000200006
4. R.F. Teófilo, J.P. Martins and M.M.C. Ferreira, "Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression", *J. Chemometr.* **23,** 32–48 (2009). https://doi.org/10.1002/cem.1192
5. J.V. Roque, W. Cardoso, L.A. Peternelli and R.F. Teófilo, "Comprehensive new approaches for variable selection using ordered predictors selection", *Anal. Chim. Acta* in press (2019). https://doi.org/10.1016/j.aca.2019.05.039
6. B.L. Legendre and D.M. Burner, "Biomass production of sugarcane cultivars and early-generation hybrids", *Biomass Bioenerg.* **8,** 55–61 (1995). https://doi.org/10.1016/0961-9534(95)00014-X
7. G.P. Meade and G.L. Spencer, *Spencer-Meade Cane Sugar Handbook*. John Wiley (1963).
8. R.W. Kennard and L.A. Stone, "Computer-aided design of experiments", *Technometrics* **11,** 137–148 (1969). https://doi.org/10.1080/00401706.1969.10490666
9. M. Barker and W. Rayens, "Partial least squares for discrimination". *J. Chemometr.* **17,** 166–173 (2003). https://doi.org/10.1002/cem.785