MATEUS TELES VITAL GONÇALVES

INTEGRATING NIR AND GENOMIC DATA FOR PREDICTING FIBER AND SUCROSE CONTENT IN SUGARCANE

Dissertation submitted to the Universidade Federal de Viçosa, in partial fulfillment of the requirements of the Genetic and Breeding Graduate Program for the degree of *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRAZIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa – Câmpus Viçosa

T
G635i
2019

Gonçalves, Mateus Teles Vital, 1993-

Integrating NIR and genomic data for predecting fiber and sucrose content in sugarcane / Mateus Teles Vital Gonçalves. – Viçosa, MG, 2019.

vii, 64f. : il. (algumas color.) ; 29 cm.

Texto em inglês.

Orientador: Luiz Alexandre Peternelli.

Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Espectroscopia de infravermelho. 2. Plantas - Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22 ed. 543.57

MATEUS TELES VITAL GONÇALVES

INTEGRATING NIR AND GENOMIC DATA FOR PREDICTING FIBER AND SUCROSE CONTENT IN SUGARCANE

Dissertation submitted to the Universidade Federal de Viçosa, in partial fulfilment of the requirements of the Genetic and Breeding Graduate Program for the degree of *Magister Scientiae*.

APPROVED: July 26, 2019.

Assent:

Mateus Teles Vital Gonçalves
Author

Luiz Alexandre Peternelli
Advisor

quando nascidos, recebemos da
sociedade um roteiro de conduta
que nos diz como fazer, pensar
e sentir. Ao seguir o roteiro,
morremos pouco a pouco

luck and misfortune, beginning with the accident of our birth in a particular class, nation, and community, shape much of what happens to us

we would be almost nothing
if we did not fight against
the consequences of this
faith and recognize in
ourselves the unresigned
and uncontainable spirits
we all really are

ROBERTO MANGABEIRA UNGER

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my parents Wilma and Gilberto for their unconditional love and for the motivation they gave me during this time. Admiration and gratitude are only a few of the many feelings I have towards you. I also would like to thank my dear brothers Gustavo, Marcelo and Vinicius for all the playful times we always have and their ever kindness feelings. To all cousins, uncles, aunts, and members of my family for the love and the good memories you provide.

I would like to acknowledge my advisor Luiz Alexandre Peternelli not only for the mentorship but also for the friendship and confidence he has pinned on me. For accepting me as a student and for his promptness for all matters and issues during this time. To my co-advisors Reinaldo Francisco Teófilo for having introduced me to the Chemometrics field and the timely suggestions. To my other co-advisor Márcio Henrique Pereira Barbosa for all the support and valuable pieces of advice during this research. To Pedro Marcus Pereira Vidigal for the immense contribution with data analysis. To Paulo Mafra de Almeida Costa and Gota Morota for accept to participate as committee members of my dissertation defense and for their contribution. To all professors, I had through my academic life. You inspired and sowed the interest to explore the uncertainties of life in me.

I want to thank all former and current members of the Multivariate Chemistry Analyses Laboratory for all end of year celebrations I was invited for. Those churrascos were great. A special thanks to Jussara for her patience to answer my questions whenever I needed and also for her care and companionship. To the good and old folks from Belo Horizonte: Gabriel, Carlos, Fred, Pedro, Bernardo, and others that would always welcome me whenever I went to my parents' house. You're my soulmates and gave me much emotional support during this time. To the uncountable friends and good people, I had the opportunity to get to know as an undergraduate student: Daniel, Rafael, Danilo, Marcus Junior, Saulo, Lucas and many others of Agronomia 2011. Republica Bola de pêlo (my first home in Viçosa), Republica Mãe Joana for all the good memories. To the friends I made during the time I lived in Budapest: Erick, Eszter, Maria, Flavia, Juliana, Lucas, Attila, Andrea, Nello, Antonio, Matheus, and many

others. I wouldn’t be able to go through this period without the learnings I got from you. I will remember of you all with great fondness.

I wish to acknowledge the public and private institutions for providing not only my scholarship but also resources and the infrastructure that made possible this work: Capes, CNPq, FAPEMIG, and Ridesa. A special thanks for all the staff of the PMGCA and Oratórios for the collection of experimental data and for the good laughs. To the graduate and undergraduate students of the LAPEA for the support. To all staff and colleagues from the Genetics and Plant breeding Graduate program. Lastly, I also would like to thank all the people of the knowledge economy for their contribution sharing tips and science on the internet, without which I would not have access to much information used when I was writing this text.

Viçosa, 2019.

TABLE OF CONTENTS

ABSTRACT

GONÇALVES, Mateus Teles Vital, M.Sc., Universidade Federal de Viçosa, July, 2019. **NIR spectroscopy and genomic selection for sugarcane breeding**. Advisor: Luiz Alexandre Peternelli. Co-advisors: Marcio Henrique Pereira Barbosa and Reinaldo Teófilo.

The main goal of this dissertation was to investigate candidate methodologies to circumvent some of the bottlenecks of the sugarcane genetic breeding program of the Universidade Federal de Viçosa (PMGCA). In chapter one, we developed regression and classification models using near-infrared spectroscopy to predict and classify sugarcane clones based on two feedstock quality parameters. The values measured by reference methods and predicted by PLS and PLS-DA models were compared. The PLS models developed had moderate accuracies. The correlation coefficients of prediction obtained were: 0.732 for fibre content and 0.665 for sucrose content. The PLS-DA models built to classify clones based on PC% showed the ideal value of 1 for sensitivity, whereas models based on FIB% showed a moderate value of 0.758. Both models exhibited similar classification errors: 0.185 and 0.195 for FIB% and PC%, respectively. These results indicate the feasibility of NIR spectroscopy coupled with multivariate analysis for the substitution of current time-consuming methods in the evaluation of large populations of sugarcane clones. In chapter two, we investigate whether the accuracy of genomic selection is improved, by combining the NIR spectra matrix to a SNP genotyping matrix into a single regression analysis. The accuracy of genomic selection models was evaluated using the Kennard-Stone algorithm and computing the correlation between the breeding values obtained using phenotypic measurements and breeding values estimated using genomic information. Combining the NIR spectroscopy information to the genomic dataset improved the correlation coefficient estimates for FIB% and PC%. The results found in this study suggest that models including NIR spectra-derived data coupled with molecular markers information resulted in higher predictive ability. Hence, this approach could be used to enhance the efficiency of selection of sugarcane clones by reducing breeding time cycles and thus, increase genetic gains at the PMGCA.

RESUMO

GONÇALVES, Mateus Teles Vital, M.Sc., Universidade Federal de Viçosa, julho, 2019. **NIR spectroscopy and genomic selection for sugarcane breeding**. Orientador: Luiz Alexandre Peternelli. Coorientadores: Marcio Henrique Pereira Barbosa e Reinaldo Teófilo.

O objetivo principal desta dissertação foi investigar novas metodologias para contornar alguns dos gargalos do programa de melhoramento genético de cana-de-açúcar da Universidade Federal de Viçosa (PMGCA). No capítulo um, desenvolvemos modelos de regressão e classificação usando espectroscopia de infravermelho próximo para prever e classificar clones de cana-de-açúcar com base em dois parâmetros de qualidade de matéria-prima. Os valores medidos por métodos de referência e previstos pelos modelos PLS e PLS-DA foram comparados. Os modelos PLS tiveram acurácias moderadas. Os coeficientes de correlação de predição obtidos foram: 0,732 para teor de fibra e 0,665 para teor de sacarose. Os modelos PLS-DA construídos para classificar clones baseados em PC% apresentaram o valor ideal de 1 para sensibilidade, enquanto os modelos baseados em FIB% apresentaram um valor moderado de 0,758. Ambos os modelos exibiram erros de classificação semelhantes: 0,185 e 0,195 para FIB% e PC%, respectivamente. Estes resultados indicam a viabilidade da espectroscopia NIR na substituição de métodos de referência adotados atualmente. No capítulo dois, investigamos se a acurácia da seleção genômica é melhorada, combinando a matriz de espectros NIR e uma matriz de SNPs em uma única análise de regressão. A acurácia dos modelos de seleção genômica foi avaliada utilizando o algoritmo de Kennard-Stone e computando a correlação entre os valores genéticos obtidos por meio de medidas fenotípicas e os valores genéticos estimados a partir da informação genômica. A combinação das informações de espectroscopia NIR e de genômica melhorou a correlação dos modelos para FIB% e PC%. Os resultados encontrados neste estudo sugerem que modelos incluindo dados derivados de espectros NIR, juntamente com informações de marcadores moleculares, resultam em maior capacidade preditiva. Assim, essa abordagem poderia ser usada para aumentar a eficiência da seleção de clones de cana-de-açúcar, reduzindo os intervalos de geração e assim possibilitar o aumento nos ganhos genéticos no PMGCA.

GENERAL INTRODUCTION

1. Sugarcane

Sugarcane (*Saccharum spp.*) is an allogamous (cross-pollinized) perennial grass. Sugarcane cultivation is made through clonal propagation. After planting, the first harvest takes place over 12 to 24 months and is referred to as plant cane. From the second harvest onwards the crops are ratoons. It is commonplace to group 6 species into the *Saccharum* genus: *S.barberi*, *S.edule, S.officinarum, S.robustum, S.sinense, and S.spontaneum*. It is assumed that the origin of the most important sugarcane species, with respect to modern cultivars, can be tracked down to Southeast Asia, in present New Guinea, Indonesia, and India (Cheavegatti-Gianotto et al., 2011). The plants cultivated by farmers and distilleries nowadays are hybrids, predominantly derived from crosses between *S.spontaneum* and *S.Officinarum* species. The combination of attributes coming from the two genitors is what provides the desired traits to modern cultivars, including high sugar yielding, tillering and ratooning ability, insect and disease resistance, drought tolerance, among others. Sugarcane is known to have a big and complex genome. This complexity arises from the interspecific hybridizations and polyploidy condition of the plant, which leads to high heterozygosity and brings along much repetitive information. Further, the cytotype (the total number of chromosomes in the cell) may vary within and between species (Cheavegatti-Gianotto et al., 2011; Hoang et al., 2015)

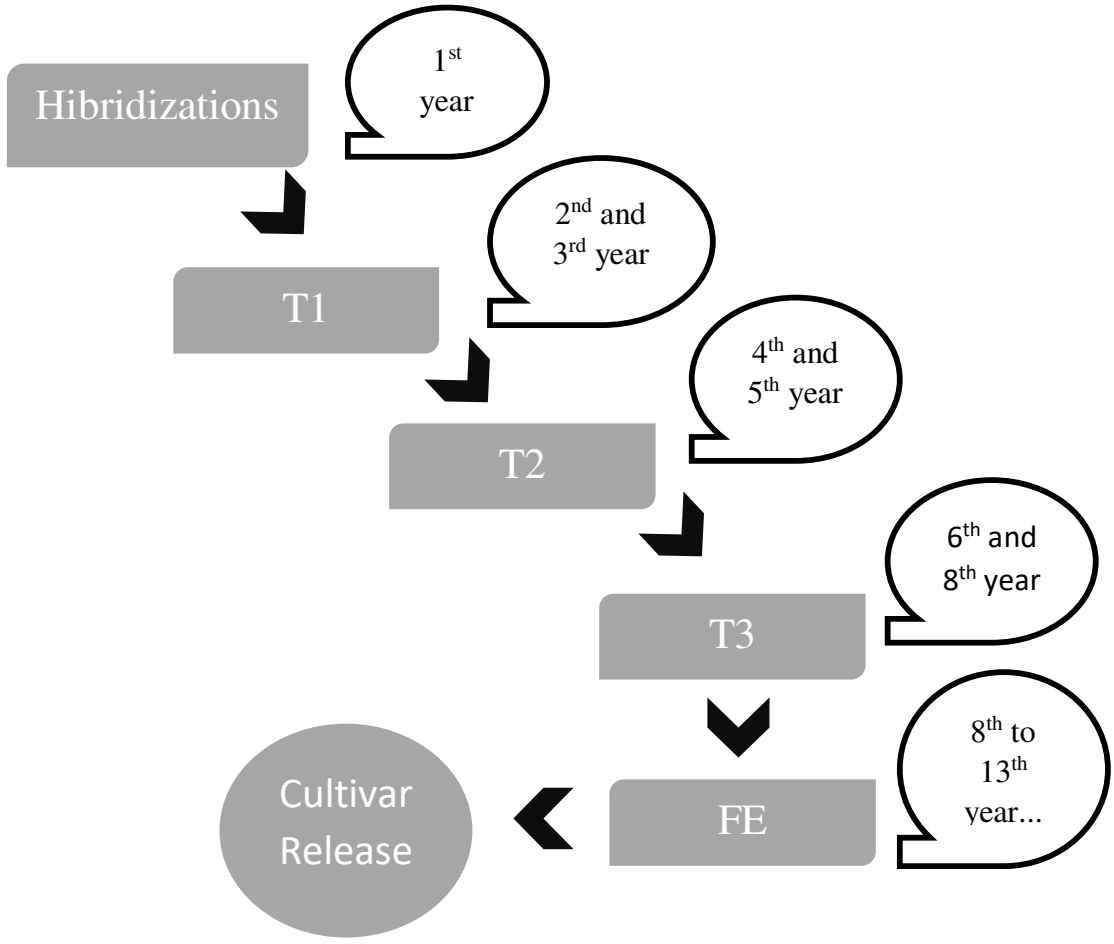
The first reports of sugarcane cultivation in Brazil date back to the colonial period when it was mainly grown for molasses and sugar production (Edel, 1969). In these days, the total area size cultivated with the crop in Brazil exceeds 8 million hectares. In addition, the Brazilian ethanol and sugar production for the crop year of 2017/2018 is estimated in 27,76 billion litres and more than 37 million tons per hectare, respectively. All of this ranks the country among the top producers and exporters of these two commodities worldwide (CONAB, 2018). Consequently, sugarcane cultivation has a significant impact on the Brazilian agribusiness sector and therefore, in the Brazilian economy.

Sugarcane is a multipurpose crop and is used as a raw material for the production of several by-products, including beverages, bio-plastic, paper, and others (Chandel et al., 2012). Nevertheless, sugarcane processing mainly yields sugar and bioethanol (whatsoever the final product, the feedstock destination along the production chain is driven by economic reasons). Moreover, sugarcane processing results in an enormous amount of a fibrous residue, known as bagasse. Sugarcane bagasse is commonly burned

at mills to produce steam and electricity, and it may also be used in the production of second-generation ethanol (Bezerra et al., 2016).

2. Sugarcane Breeding

The development of cultivars with high sucrose content has been uppermost for sugarcane breeders. However, given the growing concerns about the human activity impacts on climate change and the pursuit for alternative energy sources, sugarcane is now highlighted as a potential feedstock for biomass-based energy production (Santchurn et al., 2012). Thus, sugarcane breeding programs in Brazil and other countries are shifting their attention to sugarcane clones that are both, sucrose and high biomass yield (Da Silveira et al., 2015; Matsuoka et al., 2014). Sugarcane breeding programs vary across countries, as each project has different challenges and singularities. However, these programs typically share a common framework. The Universidade Federal de Viçosa Sugarcane Genetic Breeding Program (PMGCA) is divided into phases and is illustrated in the flowchart:



The starting line of a sugarcane breeding program is the selection of desirable genitors, pursuing genetic divergence and desirable alleles. Subsequently, the program is divided into first (T1), second (T2), third (T3) and experimental (FE) test phases. Following hybridizations, a massive number of genotypes are obtained and need to be screened to either, be advanced or dismissed from further evaluation. Therefore, being able to sort and efficiently classify these genotypes is paramount. Although the literature (Barbosa et al., 2004) indicates best linear unbiased predictor at the individual level (BLUPI) as the optimal selection approach in sugarcane breeding (where genotypic and environmental variance components are estimated), ordinary early selection of sugarcane clones at PMGCA is performed by applying mainly the mass selection approach. The mass selection approach consists of visually selecting the most suited plants or families - according to the breeder sense of judgment - in the field. Both methods have limitations, including the difficulty of collecting data of individual seedlings in the former and the absence of scientific rigor in the latter. Moreover, the current standard laboratory analytical techniques used to measure sugarcane quality parameters, e.g. polarimetry, have limitations. These methods are time-consuming, labour-intensive, and employ toxic reagents, which may jeopardize worker's health and the environment.

3. Omics Era

Nowadays, the four-letter neologism "omic" has become a ubiquitous suffix in the literature. We, the graduate students, besides scholars and professionals, when reading scientific papers, often come across words such as metabolomics, proteomics, transcriptomics, and many others. These terms are used to refer to biology systems that deal with high-dimensional molecule-level data. These biology systems aim to better understand the underlying biological phenomena of living organisms (Mangul et al., 2019). In this dissertation, two of these terms are worth to highlight: phenomics and genomics. In genetics, the phenotype can be understood as the set of visible traits that allow the observer to differentiate individuals. Therefore, phenotyping is the act of collecting phenotypes, and phenomics is the high-throughput fashion in which is done. As stated above, phenotyping sugarcane is a cumbersome task. Fortunately, breeders can increasingly count on newly developed platforms to help tackle these challenges.

Among the new available technologies NIR spectroscopy stands-out as a rapid, non-invasive, and low-cost analytical method that have been successfully applied in the pharmaceutical, food, and agricultural industry (Cozzolino, 2014). The infrared (IR)

refers to a specific region in between the visible and the microwave bands of the electromagnetic spectrum. The electromagnetic radiation is characterized by two measures, the frequency, and wavelength. The frequency is measured in hertz (Hz) and wavelengths in nanometres (nm). Often chemometricians rather choose to refer to wavelength as wavenumber, which is a conversion expressed in cm^{-1}. Throughout this text, the two terms are going to be used interchangeably. The NIR (near-infrared) locates right below the visible light and extends from 780 nm up to 2,500 nm. In Figure 1, if we move rightwards the wavelength decreases whereas the frequency of the electromagnetic radiation increases.

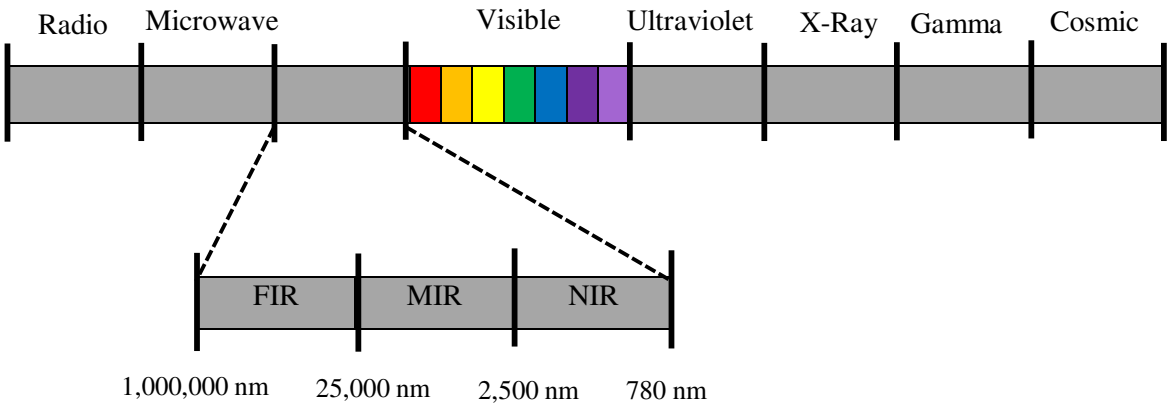


Figure 1. Illustration of the electromagnetic spectrum.

In spectroscopy, researches are interested in understanding how matter interacts with radiation to generate absorption spectra. Molecules have functional groups, and each, fundamental vibrational modes. Once excited by photon energy, derived from radiation, oscillating in the same frequency of the functional group's vibrational modes, absorption occurs. Moreover, other factors may affect the intensity of the absorption, such as the amount of change in the dipole moment of the molecule. The bigger the change, the higher is the intensity of the absortion. Therefore, for each wavelength, there is a corresponding signal associated with it. This theory is grounded in the beer law, which states that the higher the concentration of an analyte, the greater is the absorbance (Pasquini, 2003). Considering we aim to harness the NIR region to probe a sample, the resulting spectra derives mainly from the vibration modes combination, resonance, and overtones of C-H, O-H, N-H, and S-H chemical bonds.

Plant breeders still rely on phenotypic selection of segregating populations as a major means of developing elite cultivars. However, together with new developed high-

throughput phenotyping platforms, researches have increasingly been supported with new advancements in molecular biology. As a result, there has been a steady decrease in the cost of DNA sequencing technologies and an increase in the availability of molecular markers (Goodwin et al., 2016).

Meuwissen et al. (2001) were the first to propose, in a theoretical study, the usefulness of new sequencing technologies coupled with statistical methods and powerful computational tools in livestock breeding. Since then, the application of genome-wide selection (GWS) or simply genomic selection (GS) has been on the rise. The GS methodology attempts to associate DNA markers with phenotypes. The DNA markers are treated as predictor variables, which in turn have effects associated with each one of them. These effects are estimated by jointly using DNA markers information and phenotypes as a response variable. Then, it is possible to estimate genetic values in untested individuals using prediction models. Therefore, the next population to be evaluated would no longer need to be phenotyped, only genotyped (by genotyped I mean it to have DNA information collected from individual plants). This is essential in plant breeding as it may allow programs to save time and resources. The application of GS in animal breeding has proved to be efficient and to pay off (Meuwissen et al., 2013). Moreover, plant breeders have also been harnessing genomic tools to foster genetic gain of keys traits such as yield and plant resistance to biotic and abiotic factors. Furthermore, the identification of potential regions in the DNA related to agronomic traits intends to reduce breeding cycle length (Crossa et al., 2010). However, the application of genomic selection in sugarcane has been delayed when compared to other crops. Further studies investigating the potential usefulness of this approach in this crop are necessary.

4. Organization of the dissertation

This text is organised into two chapters and intends to investigate the feasibility of the application of genomic selection and NIR spectroscopy as breeding tools by the PMGCA-UFV. The chapters are presented in an article framework, containing the following sections: introduction, material and methods, results, discussion, and conclusion.

Chapter 1 provides an overview of industrial sugarcane processing and its potential as a candidate crop for biomass-based energy production. It also describes the challenges and constraints in sugarcane breeding and how NIR stands as a surrogate for the current

standard chemistry analyses adopted. The chemometrics techniques used to optimise the regression models are briefly discussed, and two regression methods are tested.

In chapter 2, the hurdles of ordinary phenotypic selection in sugarcane are presented, and new alternative technologies to overcome these constraints are discussed. We hypothesized that combining NIR spectroscopy data to a molecular marker SNP array would increase the prediction accuracy of models developed. We tested this thesis using NIR data spectra collected from ground bagasse samples of an experimental population of sugarcane clones that had been genotyped.

5. References

BARBOSA, M. H. P. *et al.* Use of REML/BLUP for the selection of sugarcane families specialized in biomass production. *Crop Breeding and Applied Biotechnology*, v. 4, n. 12, p. 218–226, 2004.

BEZERRA, T. L.; RAGAUSKAS, A. J. Review A review of sugarcane bagasse for second-generation bioethanol and biopower production. *Biofuels,Bioproducts & Biorefinering*, p. 1–14, 2016.

CHANDEL, A. K.; SILVA, S.; SINGH, O. V. Sugarcane bagasse and leaves : foreseeable biomass of biofuel and bio-products. *Journal of Chemical Technology & Biotechnology*, v. 87, n. 11, p. 11–20, 2012.

CHEAVEGATTI-GIANOTTO, A. *et al.* Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. *Tropical Plant Biology*, v. 4, n. 1, p. 62–89, 2011.

CONAB (2018) COMPAHIA NACIONAL DE ABASTECIMENTO. Cana-de-açúcar. *Acompanhamento da safra brasileira de cana-de-açúcar. Terceiro Levantamento - Safra 2018/9, Brasilia*, 2018. , p. 71.

DA SILVEIRA, L. C. I. *et al.* Selection strategy in families of energy cane based on biomass production and quality traits. *Euphytica*, n. 1, p. 1–13, 2015.

EDEL, M. The Brazilian Sugar Cycle of the Seventeenth Century and the Rise of West Indian. *Caribbean Studies*, v. 9, n. 1, p. 24–44, 1969.

GIANOLA, D. *et al.* Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers ´. *Genetics*, v. 186, n. 10, p. 713–724, 2010.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, v. 17, n. 6, p. 333–351, 2016. Disponível em: <http://dx.doi.org/10.1038/nrg.2016.49>.

HOANG, N. V. *et al.* Potential for Genetic Improvement of Sugarcane as a Source of Biomass for Biofuels. *Frontiers in Bioengineering and Biotechnology*, v. 3, n. November, p. 1–15, 2015.

MANGUL, S. *et al.* Systematic benchmarking of omics computational tools. *Nature Communications*, v. 10, n. 1393, p. 1–11, 2019. Disponível em: <http://dx.doi.org/10.1038/s41467-019-09406-4>.

MATSUOKA, S. *et al.* Energy Cane : Its Concept, Development, Characteristics, and Prospects. *Advances in Botany*, v. 2014, p. 13, 2014.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, v. 157, p. 1819–1829, 2001.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Accelerating Improvement of Livestock with Genomic Selection. *Annual Review of Animal Biosciences*, v. 1, p. 221–237, 2013.

PASQUINI, C. Near-infrared spectroscopy: Fundamentals, practical aspects, and analytical applications. *Journal of the Brazilian Chemical Society*, v. 14, n. 2, p. 198–219, 2003.

SANTCHURN, D. *et al.* From sugar industry to cane industry: Investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. *Euphytica*, v. 185, n. 3, p. 543–558, 2012.

TAYLOR, P.; COZZOLINO, D. Use of Infrared Spectroscopy for In- Field Measurement and Phenotyping of Plant Properties : Instrumentation , Data Analysis , and Examples. *Applied Spectroscopy Reviews*, v. 49, n. 7, p. 37–41, 2014.

CHAPTER 1

SCREENING SUGARCANE CLONES FOR FEEDSTOCK QUALITY TRAITS WITH NIR SPECTROSCOPY

1. Introduction

Climate change concerns linked to greenhouse gas (GHG) emissions and finite oil reservoirs have been endorsing the debate over the replacement of fossil fuels by alternative renewable energy sources (Popp et al., 2014; Sims, 2004). Hence, bioenergy crops are drawing the attention of plant science researchers (Das et al., 2017; Fernández-coppel et al., 2018; Moosavi et al., 2018). Bioenergy crops are plants used for energy production purposes, e.g., steam, electricity, fuel, and heat (Sang, 2011; Somerville et al., 2010). However, several concerns have been raised about real benefits and the sustainability of energy crops, notably those referred to as a source of first-generation feedstock (FGF). Contrary to second-generation feedstock, mainly composed of lignocellulosic biomass, FGF typically consists of crop-derived readily fermentable carbohydrates (Allwright & Taylor, 2016; Surendra et al., 2018). The main controversy over FGF is related to the food *versus* fuel fray for land usage. That is to say, whether bioenergy crops may compete with food crops for arable land, ultimately leading to a rise in food prices. Furthermore, the eventual deforestation increase lead by the expansion of energy crops cultivation to attend the demand for biofuels, thereby negatively impacting biodiversity and increasing GHG emissions (Bordonal et al., 2018; Laurance, 2007; Popp et al., 2014). Therefore, among several promising bioenergy crops, sugarcane is noteworthy (Carvalho-Netto et al., 2014; Waclawovsky et al., 2010).

Sugarcane (*Saccharum sp.)* is clonally propagated and is cultivated in tropical and subtropical regions. As a C_4 plant, sugarcane has a remarkable ability to utilize water, atmospheric CO_2, and sunlight to produce chemical energy in the form of polysaccharides, when compared to C_3 plants. As a result, it presents potential high biomass yields (Byrt et al., 2011; Henry, 2010). The Brazilian sugarcane national average yield for the crop year of 2019 is estimated at 71.3 Mg ha^{-1} (CONAB, 2018). However, these results are still far from the theoretical attainable yield of the crop (148 Mg ha^{-1}) (Carvalho-Netto et al., 2014). Moreover, more resilient and less input required sugarcane cultivars can occupy marginal and degraded land (Carvalho-Netto et al., 2014; Matsuoka et al., 2014). Indeed, virtually all sugarcane expansion in Brazil has been taking place on previously pastureland (Adami et al., 2012). In addition, most of Brazilian sugarcane mills and distilleries are located in the south-eastern region, distant from the Amazon rain forest (Cheavegatti-Gianotto et al., 2011; Macedo et al., 2008). Therefore, sugarcane would neither compete with food crops for land usage nor increase deforestation (Coelho et al.,

2006; Matsuoka et al., 2009). Lastly, environmental and conservational policies will likely ensure in the following years the sustainability of sugarcane industry in Brazil (Bordonal et al., 2018; Lewandowski et al., 2006; Nepstad et al., 2014).

Sugarcane is mostly used as raw material for sugar and first-generation bioethanol (FGB) production. Moreover, sugarcane harvest and milling process for sugar and FGB generate a considerable amount of remainders, the residual biomass (Leal et al. 2013) The residual biomass consists mainly of straw, green tops - both extracted during mechanized harvest - and bagasse. Bagasse is a non-food vegetal fibrous material obtained after stalks are crushed for juice extraction. Sugarcane bagasse major chemical constituents are cellulose, hemicellulose, and lignin (Bezerra & Ragauskas, 2016). The residual bagasse has typically two destinations at Brazilian biorefineries. It can be conveyed to mill incineration facilities to produce steam and bioelectricity, thus providing energy and attending plant industrial consumption requirements, or incorporating the energy surplus into the electric grid. Alternatively, the bagasse residuals may be used as feedstock for second-generation bioethanol production (SGB) (Bezerra et al., 2016; Carpio et al., 2019; Dias et al., 2011).

The exploitation of sugarcane bagasse as a lignocellulosic biomass for biofuel conversion still faces major challenges and is not yet economically attractive (Ali et al., 2016; de Souza et al., 2013, 2016; Wang et al., 2016). However, forecasts suggest that scientific research in genetics, agronomic management, better biotechnology, and industrial practices will gradually make SGB a feasible option (Arruda, 2011; de Souza et al., 2013; Dias et al., 2012; Lopes et al., 2016). In conclusion, sugarcane is a promising energy crop for SGB and bioelectricity production that can ultimately improve the economic viability of sugarcane biorefinaries while meeting sustainable goals (Kim et al., 2010; Marin et al., 2016).

Sugarcane breeding has traditionally been focused and guided towards the increase of sugar production (Jackson, 2005). However, a renewed interest has been diverting the attention of breeders to biomass improvement (Matsuoka et al., 2014). Thus, in order to meet future demands of the sugarcane industry productive chain and to fully harness the crop's potential, new cultivars need to be breed (Arruda, 2011; Barbosa et al., 2012; Silva et al., 2017). The biomass feedstock composition is a paramount aspect to consider in the energy conversion process, either for SGB or bioelectricity production (Surendra et al., 2018). Therefore, breeding sugarcane for biomass quality and sucrose

yield involves the assessment of key traits. For instance, information regarding sugarcane clone's fibre content (FIB%) is useful to aid breeders during the selection favouring the increase in lignocellulosic biomass yields. Additionally, apparent sucrose (PC%) measures are valuable to estimate production costs as it is a function of total recoverable sucrose, an important industrial parameter used to estimate sugar and bioethanol yields (Hoang et al., 2017). In this view, from the standpoint of Brazilian biorefineries, sugarcane cultivars with higher fibre rates and yet with reasonable amounts of sucrose are highly desirable and could bring more economical and environmental sustainability to the industry. Ideally, if bagasse destination is the cogeneration system in which the feedstock is burned to produce bioelectricity, bagasse with high lignin rates is desirable. On the other hand, if the feedstock destination is SGB production, lower lignin rates are preferable (Mendu et al., 2012; Sticklen, 2006).

From the perspective of sugarcane breeders, the *Saccharum* genus has high genetic diversity and allow the selection for a wide range of desirable traits, including higher yield and better biomass quality (da Silveira et al., 2015; Legendre & Burner, 1995; Santchurn et al., 2014). However, it can take up to 15 years until the release of new cultivars by sugarcane breeding programs (Barbosa et al., 2012; Kandel et al., 2018; Matsuoka et al., 2009). One of the main hurdles in this process is the need to screen the massive number of clones obtained after hybridizations are done in each breeding cycle. Consequently, screening sugarcane clones at the initial phases of a breeding program is a high-cost, time-consuming, and labour-intensive task. In addition, the reference chemical analyses to determine industrial quality parameters involves the use of toxic and harmful reagents such as lead acetate that may end up contaminating laboratory users and render environment impacts if not properly discarded.

NIR (near-infrared) spectroscopy is a likely replacement technique for the standard wet chemistry analyses currently adopted. This analytical technique has several attractive features, namely, velocity (less than 1 minute to acquire a spectrum), non-destructiveness, and operating easiness. In addition, it has environmental friendly aspects, since it does not generate any residual. The interaction between matter, i.e., the sample being probed, and its chemical species with near-infrared region radiation of the electromagnetic spectrum is what underlies NIR spectroscopy. The technique exploits the potential vibrational energy present in molecule bonds. Once the output of a NIR instrument and standard reference values are obtained, it is possible to build prediction or

classification models by correlating individual plant phenotypes with spectra wavelengths (Porto et al., 2019; Purcell et al., 2009; Valderrama et al., 2007). NIR modelling involves the selection of a representative set of samples, followed by the determination of chemical species of interest by a reference well-accepted analytical method. Thereafter, the collection of spectra using a spectrometer and the application of computational and statistical procedures to optimise and validate models developed (Pasquini, 2003).

The literature comprising the application of NIR spectroscopy to quantify feedstock chemical quality parameters in sugarcane is vast. Valderrama (2007) and Nawi (2014) successfully attempted to determine sugarcane samples' soluble solids (Brix%) and reducing sugars (glucose and fructose) using NIR. In recent studies, Hoang (2017) and Assis (2017) developed NIR-based models to predict sugarcane fibre content and its lignocellulosic components using ground bagasse samples. Phuphaphud (2019) assessed the feasibility of NIR spectroscopy to determine fibre content in sugarcane by directly scanning the stalks surface. However, we believe that there are no studies reporting the use of bagasse samples to simultaneously predict apparent sucrose (PC%) and fibre content (FIB%). Aiming to contribute to the Universidade Federal de Viçosa Sugarcane Genetic Breeding Program (PMGCA), we conducted investigations applying NIR spectroscopy to profile sugarcane clone's tissue chemical constituents. The goal of this study is to assess whether it is possible (1) to build robust quantitative and qualitative NIR- based models to predict sugarcane clone's feedstock quality parameters, namely FIB%, and PC%. Specific goals are (2) to rank sugarcane clones based on their genetic merit using NIR spectroscopy; (3) to compare partial least squares (PLS) and partial least squares for linear discrimination (PLS-DA) regarding their usefulness as breeding tools at the PMGCA.

2. Material and methods

2.1. Plant material

In this study, we evaluated three hundred and ninety-seven clones derived from an originally seedling population of 98 half-sib families. At the early stages of sugarcane breeding programs, selection of best families coupled with individual seedling selection within the best families is a standard procedure (Stringer et al., 2011). The seedling population - in which each plant is a single genotype - was the result of crosses made at the Serra do Ouro Flowering and Breeding Station, municipality of Murici, Alagoas State, Brazil (09°13' S, 35°50' W, 450 m altitude). After processing, seeds were sent to the Sugarcane Genetic Breeding Research Station (CECA) of the Universidade Federal de Viçosa, municipality of Oratórios, Minas Gerais State, Brazil (20°25' S, 42°48' W, 494 m altitude) and germinated in a nursery house. Subsequently, seedlings obtained from each family were transplanted to the field and evaluated in first (plant cane) and second (ratoon) crops based on desirable traits in first (T1) and second (T2) clonal trial (Barbosa et al., 2012).

2.2. Experimental design

An augmented block design was installed in May 2016, at CECA municipality of Oratórios, Minas Gerais State, Brazil (20°25' S, 42°48' W, 494 m altitude). The checks – released cultivars RB867515, RB966928, and RB92579 – were included once in each block, and regular treatments were arranged in 21 blocks (Federer, 1961). In total, 18 blocks contained 24 plots; one block contained 16 plots, and two blocks contained 12 plots. The experimental plots consisted of double-row 3 m long furrows x 1.4 m between rows and clones were cultivated following standard commercial agronomic protocols regarding fertilization, weed control and pest management (Leite et al., 2006). The whole experiment area was encompassed by buffer rows of a released commercial cultivar (Figure 1).

Site: PMGCA experimental station, Oratorios municipality, Brazil.
Plantation date: May 5th, 2016
Design: Augmented Blocks, 3 commom factors, 409 regular factors

Common treatments

RB867515	P1
RB966928	P2
RB92579	P3

3m BUFFER FURROW

BUFFER FURROW

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	F26	F27	F28	F29	F30	F31	F32	F33	F34	F35	F36	F37	F38	F39	F40
1	13	22	34	43	55	64	P1	85	97	106	P3	127	139	148	160	169	181	190	P1	211	223	232	P3	253	265	274	286	295	307	316	P1	337	349	358	P3	379	388	397	406
2	P1	23	35	44	56	65	76	86	98	107	118	128	P1	149	161	170	182	191	202	212	224	233	244	254	P1	275	287	296	308	317	328	338	350	359	370	P1	389	398	407
3	14	24	P2	45	57	66	77	87	99	108	119	129	140	150	P2	171	183	192	203	213	225	234	245	255	266	276	P2	297	309	318	329	339	351	360	371	380	P2	399	408
4	15	25	36	46	P3	67	78	88	P2	109	120	130	141	151	162	172	P3	193	204	214	P2	235	246	256	267	277	288	298	P3	319	330	340	P2	361	372	381	390	P3	409
5	16	26	37	47	58	68	P1	89	100	110	P3	131	142	152	163	173	184	194	P1	215	226	236	P3	257	268	278	289	299	310	320	P1	341	352	362	P3	382	391	400	P1
6	P1	27	38	48	59	69	79	90	101	111	121	132	P1	153	164	174	185	195	205	216	227	237	247	258	P1	279	290	300	311	321	331	342	353	363	373	P1	392	401	P1
7	17	28	P2	49	60	70	80	91	102	112	122	133	143	154	P2	175	186	196	206	217	228	238	248	259	269	280	P2	301	312	322	332	343	354	364	374	383	P2	402	P1
8	18	29	39	50	P3	71	81	92	P2	113	123	134	144	155	165	176	P3	197	207	218	P2	239	249	260	270	281	291	302	P3	323	333	344	P2	365	375	384	393	P3	P1
9	19	30	40	51	61	72	P1	93	103	114	P3	135	145	156	166	177	187	198	P1	219	229	240	P3	261	271	282	292	303	313	324	P1	345	355	366	P3	385	394	403	P1
10	P1	31	41	52	62	73	82	94	104	115	124	136	P1	157	167	178	188	199	208	220	230	241	250	262	P1	283	293	304	314	325	334	346	356	367	376	P1	395	404	P1
11	20	32	P2	53	63	74	83	95	105	116	125	137	146	158	P2	179	189	200	209	221	231	242	251	263	272	284	P2	305	315	326	335	347	357	368	377	386	P2	405	P1
12	21	33	42	54	P3	75	84	96	P2	117	126	138	147	159	168	180	P3	201	210	222	P2	243	252	264	273	285	294	306	P3	327	336	348	P2	369	378	387	396	P3	P1

3m BUFFER FURROW

BUFFER FURROW

BLOCK 1	BLOCK 4	BLOCK 7	BLOCK 10	BLOCK 13	BLOCK 16	BLOCK 19
BLOCK 2	BLOCK 5	BLOCK 8	BLOCK 11	BLOCK 14	BLOCK 17	BLOCK 20
BLOCK 3	BLOCK 6	BLOCK 9	BLOCK 12	BLOCK 15	BLOCK 18	BLOCK 21

Figure 1. Map of the experimental field.

2.3. Phenotypic data and reference analysis

The clones were evaluated in first crop (12 months after planting) and second crop (26 months after planting). In this study, we only used data of the second crop. The methods employed to estimate percentage of soluble solids in cane juice (BRIX %), sucrose in juice (POL%), apparent sucrose in cane (PC%), and fibre content (FIB%), followed recommendations of the CONSECANA manual (Consecana, 2006), which is routinely applied in sugarcane breeding programs in Brazil. At Brazilian biorefineries and distilleries, these are standard quality parameters used to classify sugarcane feedstock and determine prices. Harvest and quality analyses were performed in July 2018. Aiming to work with a representative set of the samples, ten randomly selected stalks from each of the double-row plots were cut at ground level with a machete. Green tops, clinging leaves and leaf sheaths were removed before stalks were bundled and weighted using a dynamometer. After, the ten randomly selected clones in each plot were shredded using a stationary forage chopper machine (EN-6500 Model, Nogueira & Brait industries, Itapira, Brazil). One subsample of 500 g from the shredded stalks was collected and pressed with a hydraulic press (PL011 Model, Dedini, Inc., Piracicaba, Brazil) at 250 kgf/cm^2 (24.5 MPa) for 1 minute. After pressing, juice and remainder fibre cake were collected and took to the laboratory. The juice was analysed for BRIX% using a refractometer (HI96801 Model, Hanna® instruments, Woonsocket, USA) and for POL% by polarimetry using a saccharimeter (SDA2500 Model, Acatec, Brazil) after clarifying the solution with lead acetate. The remainder fibre cake was weighed (*WC*) and used to compute fibre content:

$$FIB\% = 0{,}08 \times WC + 0{,}876$$

The apparent percentage of sucrose in sugarcane (PC %) was computed on the basis of POL% and FIB% as follows:

$$PC\% = POL\% \times (1 - 0{,}01 \times FIB\%) \times C$$

where *C* is the coefficient to convert sucrose of juice into sucrose of cane and is calculated using the formula $C = 1{,}0313 - 0{,}00575 \times FIB\%$. The final values were all expressed on the total fresh biomass basis (500g of shredded stalks).

2.4. Sample preparation and NIR spectra acquisition

Another subsample of 100 g from the shredded stalks was collected and immediately took to dry in a forced-air circulating oven at 50 °C for 24 h or until a constant mass was reached. After, to obtain a homogenous particle size, dried samples were ground using a mill with a 0.4 mm bottom sieve plate, packaged in a plastic zip bag and stored. Sample preparation process took two months in total (processing of each sample took approximately 5 minutes). The NIR spectra of samples were measured in indoor-conditions at room temperature of 21 ºC. The instrument used was a Fourier transform near infrared (FT-NIR) spectrometer (Antaris™ II Model, Thermo Scientific Inc., USA) with an integrating sphere diffuse reflectance module operating with the TQ Analysis software. Instrument operating conditions were set as follows: 4 cm^{-1} resolution in an investigated wavenumber range of 10000 - 4000 cm^{-1} and reflectance mode as log (1/R), where R is the measured reflectance. Samples were placed into a powder sampling cup accessory and arranged onto the instrument window. At each scan, the accessory was moved as to cover different positions of the sample, near-infrared six positions. A single scan measure was the average result of 32 scans, thus for each sample a total of 192 scans were made and then averaged, representing the final spectrum.

2.5. Statistical Analysis

Reliable reference values are key factors to attain good modelling. Inaccurate estimations heavily impact the usefulness and ability of NIR based-models to infer external sample properties. Thus, accurate estimations of the sugarcane genotypes genetic merit to be ranked is strongly desired. According to Barbosa et al (2004), best linear unbiased predictor at individual level (BLUPI) is the optimal selection approach in sugarcane breeding. Therefore, in order to correctly rank clones in respect to their genetic merit, variance components and the genotypic values ($\mu + \hat{g}$) for each trait were estimated by mixed model equation means using the SELEGEN-REML/BLUP software (Resende, 2016), based on the statistical model:

$$y = Xa + Zg + Wb + e \ ,$$

where y is the phenotypic observations vector; a is the vector of fixed effects (overall mean or intercept); g is the vector of random genotypic effects ($g \sim N(0, I\sigma^2)$).; b is the vector of random block effects and e is the random residue vector effects ($e \sim N(0,$

σ^2)). The capital letters X , Z and W represent the incidence matrices of the respective random and fixed effects (Henderson, 1975).

2.6. Chemometrics

Chemometrics is a branch of analytical chemistry that deals with analytic instrument-derived data using mathematical, statistical, and computational tools to do so. The emergence of chemometrics was possible due to progress made in instrumentation and computational power, opening the door for a steeply increase in storage capacity, not to mention data collection and processing capabilities (Brereton, 2014). A NIR instrument output is a two-dimensional matrix, where rows represent samples (observations) and columns predictor variables (wavelengths). NIR data is a high-dimensional dataset, as for each sample, thousands of predictor variables can be collected (Brereton, 2000). Hence, data matrices are ill-conditioned and exhibit a large degree of multicollinearity, posing several challenges and restrictions to the statistical analysis (Ildiko et al., 1993). Consequently, the use of chemometric techniques is what allows the translation of a NIR spectrum into useful information. The process of using a NIR instrument signal to build regression and classification models is characterized as a multivariate analysis task and it takes the form of the standard multiple linear regression model:

$$Y = \sum^{p} \beta_p X_p + \varepsilon ,$$

where **Y** is the response variable, e.g., FIB% concentration; β represents the parameters or regression coefficients; p is the total number of variables; **X** is the NIR data matrix, and ε is the random error term associated with each observation.

NIR data analysis involves three major steps: preprocessing, multivariate calibration, and validation (Blanco & Villarroya, 2002; Ferreira et al., 1999). The prepossessing step involves the removal of systematic variation of data matrix as the signal of a NIR instrument contains chemical information of the sample being probed along with attached noise, i.e., unwanted variation not related with the response variable (Wold et al., 1998). In this view, a NIR data output is initially referred to as raw because it has not gone through any mathematical transformation. However, often, that is the case, since multiple sources of factors may negatively influence modelling quality. For instance, physical aspects like particle size and distribution might lead to deviations in

the infra-red light beam optical path and light scattering, which frequently occurs in solid samples such as sugarcane ground bagasse (Sabatier et al., 2011). Moreover, instrument components variations, e.g. detector and light source, temperature, sample concentration variation, and other chemical interferers may also affect the results (Rinnan et al., 2009).

After preprocessing, the next step consists of splitting the data in training and test sets. The training set will be used to adjust models and the test set to validate candidate developed models. From this viewpoint, it is imperative to select an adequate training set containing all data variability to obtain robust predictions of future analysis. Therefore, in this study, samples were split into two subsets applying the Kennard-Stone algorithm (KS) (Kennard & Stone, 1969). Kennard-Stone algorithm is a step-by-step procedure that aims to maximize Euclidean distance between pairs of vectors of the **X** data matrix and a candidate sample, thereby providing an even assignment of samples to both training and test subsets (Galvão et al., 2005; Zhang et al., 2017).

In order to get more information about candidate models, resampling is employed in multivariate calibration of NIR data. This strategy is worthwhile both for quantitative and qualitative purposes since it might yield improved prediction accuracy, as it provides estimates of the test error rates using only observations of the training set. Therefore, optimizing the process of choosing the number of principal components (or latent variables). In cross-validation, a subset of samples are repeatedly drawn out from the training set pool, left aside and then used to validate regression lines, fitted in the training set samples (James et al., 2007). The root mean squared error of cross-validation (RMSECV) is computed for each prediction of the removed subset as follows:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}},$$

where, y_i is the measured value for the *i*th sample; $\hat{y}_i$ is the predicted value for the *i*th sample and N is the total number of samples (Pasquini, 2003). NIR-derived matrices are regarded as noisy provided that it frequently contains a high level of undesired variation. Moreover, the high dimensional setting adds up great amounts of redundant information, resulting in collinearity among vectors as the number of variables is much larger than the number of observations. All these features hamper the multivariate calibration step.

Therefore, FIB% and PC% models were developed using PLS (partial least squares) and PLS-DA (partial least squares discriminant analysis) regression.

PLS is arguably the most employed multivariate regression method in the chemometrics field (Pasquini, 2018). PLS regression intends to search and maximize a covariance framework between the response (reference values) and predictor variables (NIR spectra wavelengths) presuming a linear relationship between spectra and the chemical species concentration of interest. The PLS approach is categorized as a dimension reduction method as overcomes the high–dimensional issue by transforming the predictive variable set, making previous related variables to be uncorrelated (Roque et al., 2017; Wold et al., 2001). In contrast to other dimension reduction methods such as principal components regression (PCR), PLS regression is considered to be a supervised regression method since it incorporates the response variable information when identifying the directions of the principal components in the multidimensional space.

There are several criteria when answering the question of whether a model is considered to be good or not. In this study, we adopted measures regarding model accuracy. The residuals ε were used to calculate the root mean squared error of prediction for the validation step (RMSEP). In the validation step, an external set of samples not used to adjust the model are predicted, and this measure is used to represent the prediction accuracy:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}},$$

where, y_i is the measured value for the *i*th sample; $\hat{y}_i$ is the predicted value for the *i*th sample, and N is the total number of samples (Pasquini, 2003).. Another estimator of the quality of the model was the correlation coefficient that measures how strong the fitted regression line is related to the response variable (Kjeldahl et al., 2010). It can be computed as follows:

$$rp = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}},$$

where, y_i is the measured value for the *i*th sample; $\hat{y}_i$ is the predicted value for the *i*th sample, $\overline{y}$ is the mean value for the *ith* sample and $\overline{\hat{y}}$ is the mean estimated value for the *ith* sample (Roque et al., 2017). Moreover, the relative error rate was also used to evaluate models. The relative error of each sample is given by:

$$Er\% = \frac{(y_i - \hat{y}_i)}{y_i} 100,$$

where y_i is the measured value for the *i*th sample; $\hat{y}_i$ is the predicted value for the *i*th sample. The PLS-DA is an extension of PLS regression and is a supervised pattern-recognition method. It is rather an algorithm that incorporates linear discriminant analysis (LDA) and PLS into a single procedure (Barker et al., 2003; Brereton et al., 2014). PLS-DA is a classification approach, and it is used when the response variable is qualitative. In this study, we only had a quantitative data type. Thus we transformed it into a dummy variable coded as a binary response:

$$Y = \begin{cases} \textbf{1} \text{ if selected} \\ \textbf{0} \text{ if not selected} \end{cases}$$

where we only assumed to have two classes, class **1** if the response was above the mean and class **2** if the response was below the mean. The performance of classification models can be assessed in terms of classification error and classification accuracy (Lee et al., 2018). From the confusion matrix template illustrated below, the statistics that measure model quality can be computed:

	Class 1	Other than 1
Predicted as class 1	TP	FP
Predicted as other than 1	FN	TN

where, TP, TN, FP, and FN stands for true positive, true negative, false positive and false negative, respectively. From the values of the confusion matrix, the following formulas are derived:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP},$$

$$\text{Classification error} = \frac{FP + FN}{TP + TN + FP + FN}$$

The sensitivity can be interpreted as the model ability to assign a sample to the class in which it belongs, whereas the specificity is regarded as the model ability to identify samples not belonging to a given class. In both measures, the closer to 1 the better. Additionally, the classification error is the ratio of all misclassified samples to the total number of samples (Ballabio et al., 2013; Porto et al., 2019).The inclusion of many irrelevant variables, which is to say variables not associated with the response may bring great variation and diminish model performance.

Variable selection may improve model interpretability and provide a more comprehensive study of the system being assessed (Mehmood et al., 2012). Hence, we applied the ordered predictor selection (OPS) method (Teófilo et al., 2009). In short, the OPS algorithm initially obtains an informative vector from a PLS model and, provided their absolute value, vectors are sorted in descending order. Next, an initial subset of variables (window) is chosen, and a separate OLS regression is fitted. This step is repeated for all possible different combinations of *p* variables. Lastly, the best model is chosen based on the cross-validation error estimates (RMSECV; Guimarães et al., 2016; Teófilo et al., 2009). All chemometrics analyses were executed using in-house programmed codes with the assistance of Matlab (Matlab R2016a, 9.0, The MathWorks Inc., Natick, USA) and PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, USA). The OPS algorithm can be found in http://www.deq.ufv.br/chemometrics.

3. Results and discussion

Deviance analysis indicated the significance ($\rho < 0.01$) of genotypic effects, suggesting the existence of available genetic variance to be exploited for selection purposes (Table1). Variance components estimates were used to compute genetic and environmental parameters. Genetic and environmental parameters estimates allowed the inference of the genetic gain and heritability of the two traits evaluated.

Table1. Genetic and environmental parameter estimates of the 397 sugarcane clones.

Parameters	FIB%	PC%
$\hat{\sigma}_g^2$	1.36*	1.07*
$\hat{\sigma}_e^2$	0.19	2.30
h_g^2	0.95	0.65
Accuracy	0.98	0.80
Mean	13.46	15.70

$\hat{\sigma}_g^2$: genotypic variance effect; h_g^2: individual plots broad sense heritability; * significant at 1% probability by the deviance analyses; FIB%: fibre content; PC%: apparent sugar in cane.

The heritability of FIB% was higher than PC%, indicating that the eventual selection of clones for FIB% based on phenotype values would be effective. The best linear unbiased predictions (BLUPs) of genotypic effects of the best 50 clones for FIB% are represented in Table 2 and for PC% in Table 3.

Table 2. Genotypic effects of the 50 best clones evaluated for fibre content (FIB%)

Clone	µ+g	Clone	µ+g	Clone	µ+g
8	16.6774	**228**	15.3887	**34**	14.9009
91	16.4910	**400**	15.3877	**371**	14.8948
340	16.0730	**140**	15.3786	**86**	14.8800
5	16.0268	**72**	15.3656	**174**	14.8631
342	15.9095	**186**	15.3379	**366**	14.8589
344	15.8040	**305**	15.2223	**311**	14.8575
191	15.7596	**372**	15.1938	**303**	14.8442
193	15.7157	**347**	15.1666	**374**	14.8280
288	15.6770	**16**	15.1563	**398**	14.7941
165	15.6368	**380**	15.1107	**233**	14.7925
170	15.5808	**284**	14.9761	**33**	14.7917
184	15.5753	**163**	14.9686	**25**	14.7515
1	15.5516	**277**	14.9472	**274**	14.7362
57	15.4900	**98**	14.9240	**52**	14.7213
244	15.4783	**278**	14.9190	**389**	14.7150
204	15.4079	**156**	14.9090	**175**	14.7048
43	15.3933	**31**	14.9060		

µ+g: genotypic effect.

Table 3. Genotypic effects of the 50 best clones evaluated for apparent sugar in cane (PC%).

Clone	μ+g	Clone	μ+g	Clone	μ+g
134	17.4918	**147**	16.6526	**76**	16.4493
116	17.2521	**79**	16.6080	**119**	16.4493
83	17.1027	**37**	16.6049	**126**	16.4450
39	17.0997	**19**	16.6017	**131**	16.4433
133	16.9549	**264**	16.5946	**96**	16.4418
136	16.9449	**217**	16.5899	**320**	16.4275
284	16.8615	**82**	16.5753	**324**	16.4186
123	16.8399	**14**	16.5386	**366**	16.4122
130	16.8220	**23**	16.5386	**65**	16.4017
112	16.8113	**310**	16.5321	**225**	16.3967
63	16.7987	**64**	16.5256	**249**	16.3865
105	16.7850	**382**	16.5232	**301**	16.3764
29	16.7701	**102**	16.5190	**314**	16.3531
59	16.7605	**33**	16.4989	**74**	16.3529
77	16.7515	**144**	16.4782	**343**	16.3481
25	16.7070	**78**	16.4620	**319**	16.3429
69	16.6938	**129**	16.4598		

μ+g: genotypic effect.

The negative correlation results found by Hoang et al. (2017) and Santchurn et al., (2014) suggest that FIB% and PC% are opposed traits, hinting a portioning mechanism whereby the increase of one would diminish the other, and therefore selecting clones possessing high FIB% and PC% rates is not achievable. However, it is worth to note that among the best clones selected for FIB% and PC%, four clones (25, 33, 284 and 366) are present in both ranking lists. Thus, these results indicate that indeed, it is possible to obtain sugarcane clones combining both attributes, high fibre, and sugar yielding in a segregating population (Ramos et al., 2017). The results of standard chemistry analysis of the three hundred and ninety-seven (excluding checks) sugarcane samples are shown in Table 4. PC% data had a higher mean (15.08) and standard deviation (1.83) than FIB%. On a fresh biomass basis of shredded stalks, sugarcane clones sampled in this study had fibre content and apparent sucrose values ranging from 10.53-17.76% and 9.31-20.69%, respectively.

Table 4. Descriptive statistics of the standard chemistry analysis

	Mean	Standard deviation	Range (min-max)
FIB%	13.71	1.27	10.53-17.76
PC%	15.08	1.83	9.31-20.69

FIB%: fibre content; PC%: apparent sugar in cane; min: minimum; máx: maximum.

In order to divide the data into two sets, Kennard-Stone algorithm was employed, and 360 samples were allocated to the calibration set and 100 samples to the validation set Table 5. It is imperative that the calibration set, which is used to build models has a sufficient number of samples covering the whole existing variation in the dataset. Calibration and validation sets contained similar features regarding mean, standard deviation, and the range of concentration of the analyte of interest.

Table 5. Summary of the descriptive statistics of the standard chemistry analysis for fibre content (FIB%) and apparent percentage of sugar in cane (PC%) of the calibration and validation sets.

	FIB%	PC%
Calibration		
n	360	360
mean	13.69	15.16
std	1.2	1.8
range (min-máx)	(10.59-17.76)	(9.31-20.69)
Validation		
n	100	100
mean	13.58	15.33
std	1.25	1.94
range (min-máx)	(10.53-16.54)	(9.58-20.00)

n: number of samples; std: standard deviation; min: minimum; máx: maximum.

3.1. PLS results

In this section, the models developed for FIB% and PC% using the PLS regression approach are shown. The NIR spectra matrix was mathematically treated applying row-wise and column-wise preprocessing techniques. Several combinations and order of combinations were compared and the ones that yielded the best models were Savitzky-Golay smoothing (polynomial order: 2, window width: 15 points), 1st derivative

(polynomial order: 3, window width: 15 points) and mean centre for FIB%. For PC% the best combination was mean centre, Savitkzy-Golay smoothing (polynomial order: 6, window width :15 points) and 1st derivative (polynomial order: 3, window width: 25 points). The raw and treated NIR spectra of the dry ground sugarcane samples can be seen in Figure 2.

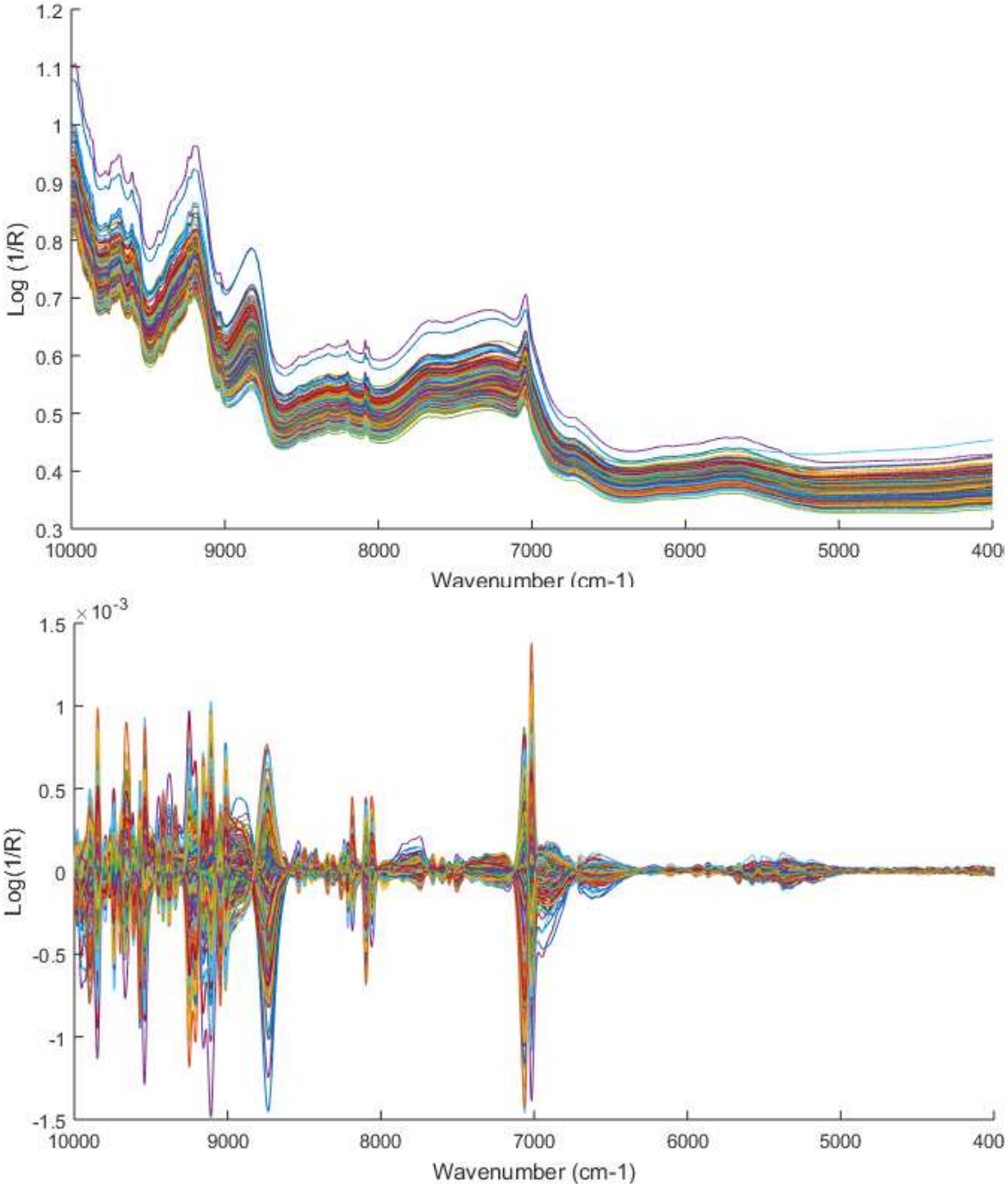


Figure 2. NIR spectra of the dry ground sugarcane samples. *Top:* raw spectra. *Bottom:* treated spectra.

In this study, we chose 10-fold cross-validation as the resampling strategy to optimize the selection of the number of latent variables to be included in the regression

fit, which in turn was determined based on the minimum values of RMSECV. To evaluate the reliability of the models, RMSEP and the relative errors were computed when the developed model was applied to predict a set of unknown samples. Table 6 shows the comparative performance of full and OPS PLS models for FIB% and PC%. Table 6 shows that the number of latent variables required to build the PLS models was 9 for FIB% and 8 for PC%.

Table 6. Performance and statistical measures of PLS models for FIB% and PC% with full and selected variables by the OPS method.

	FIB%		PC%	
	Full	OPS	Full	OPS
h	9	9	8	8
nVars	3112	185	3112	155
RMSEC	1.303	1.209	1.368	1.213
Rc	0.690	0.740	0.649	0.739
RMSECV	1.470	1.358	1.482	1.360
Rcv	0.5894	0.663	0.575	0.665
RMSEP	1.442	1.743	1.540	1.701
Rp	0.732	0.587	0.665	0.622

h: number of latent variables; nVars: number of variables; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross-validation; RMSEP: root mean square error of prediction; R_c: correlation coefficient of calibration; R_{cv}: correlation coefficient of cross-validation; Rp: correlation coefficient of prediction; FIB%: fibre content; PC: apparent sucrose in cane; FIB%: fibre content; PC%: apparent sugar in cane.

In principle, the highest the number of latent variables the highest are the chances of overfitting, that is to say, a developed model may perform well in the calibration set but poorly when predicting unknown samples in the validation step. However, sugarcane bagasse samples have a quite complex composition, thus justifying the higher number of latent variables utilized (Assis, 2017). The accuracy of models can be evaluated based on the correlation coefficient values. The full FIB% model had a higher correlation coefficient (R_p) than the full PC% model. In addition, the RMSEP of the FIB% model was smaller than PC%. The goodness of fit can be assessed by the graph of measured and predicted values (Figure 3). Although the prediction errors were relatively high, a linear pattern between measured and predicted values can be seen (Figure 3). The OPS method selected 185 variables for the FIB% data set and 155 for PC%. The correlation coefficient

of the OPS models for FIB% and PC% were smaller when compared to the values obtained using the whole set of variables (Table 6).

Additionally, the OPS models for FIB% and PC% presented an increase in the values of RMSEP. Regarding the relative error values of the full model and the OPS model, we can see that there is also no observable improvement for both traits (Figure 4). The correlation coefficient of the OPS models for FIB% and PC% were smaller when compared to the values obtained using the whole set of variables (Table 6). Moreover, the OPS models for FIB% and PC% presented an increase in the values of RMSEP. Regarding the relative error values of the full model and of the OPS model we can see that there is also no observable improvement for both traits (Figure 4). These results are in agreement with Phuphaphud et al (2019), who obtained a 0.75 correlation coefficient for fibre content prediction. However, in this study, the authors used visible-shortwave NIR spectroscopy and collected spectra directly from sugarcane stalks.

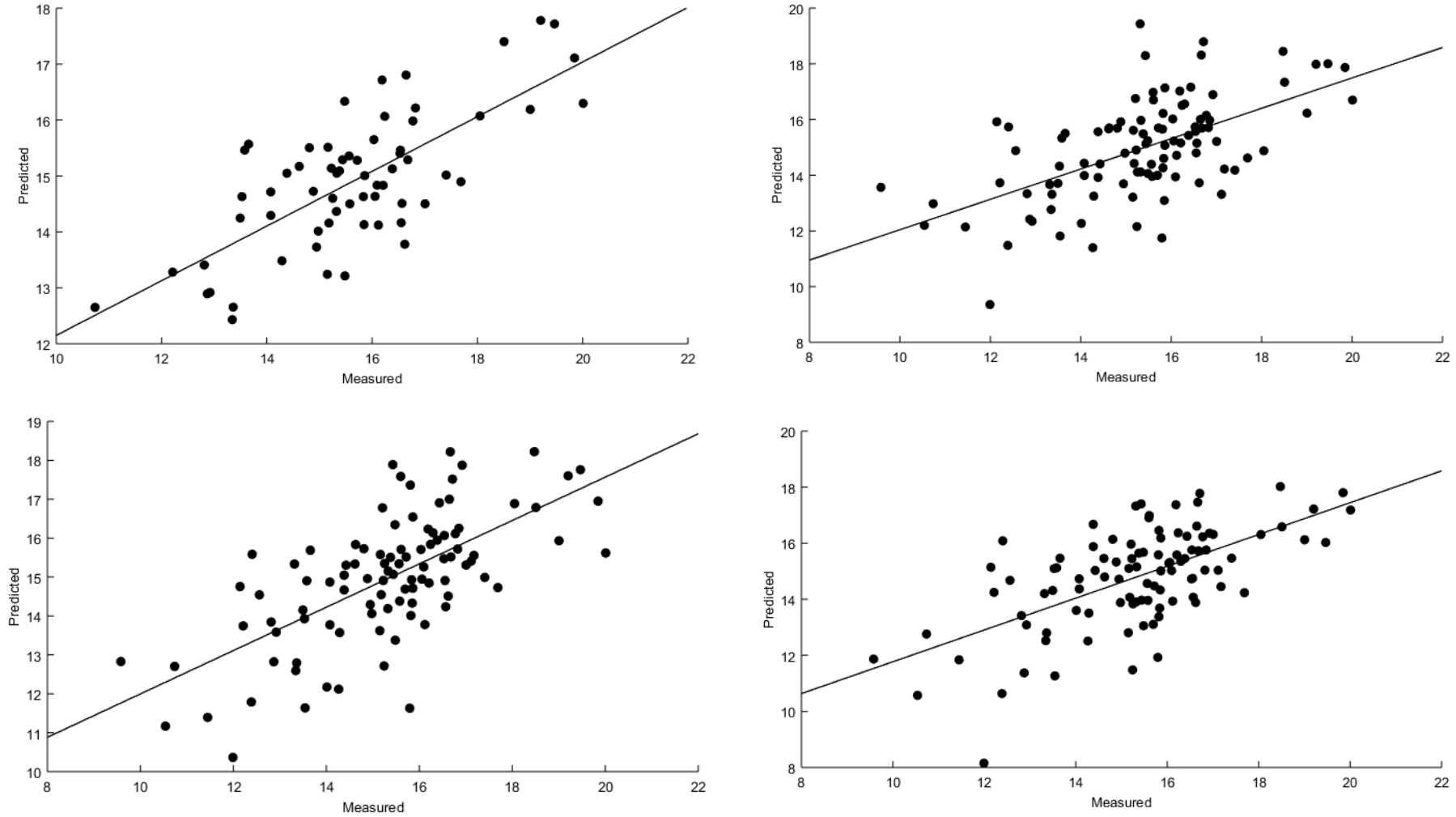


Figure 3. Predictions of FIB% and PC% for for ground bagasse sugarcane samples. *Top left panel*: full model plot of the measured versus predicted fibre content (FIB%) values. *Top right panel*: OPS model plot of the measured versus predicted fibre content (FIB%) values. *Bottom left panel*: full model plot of the measured versus predicted apparent sucrose in cane (PC%) values. *Bottom right panel*: OPS model plot of the measured versus predicted apparent sucrose in cane (PC%) values.

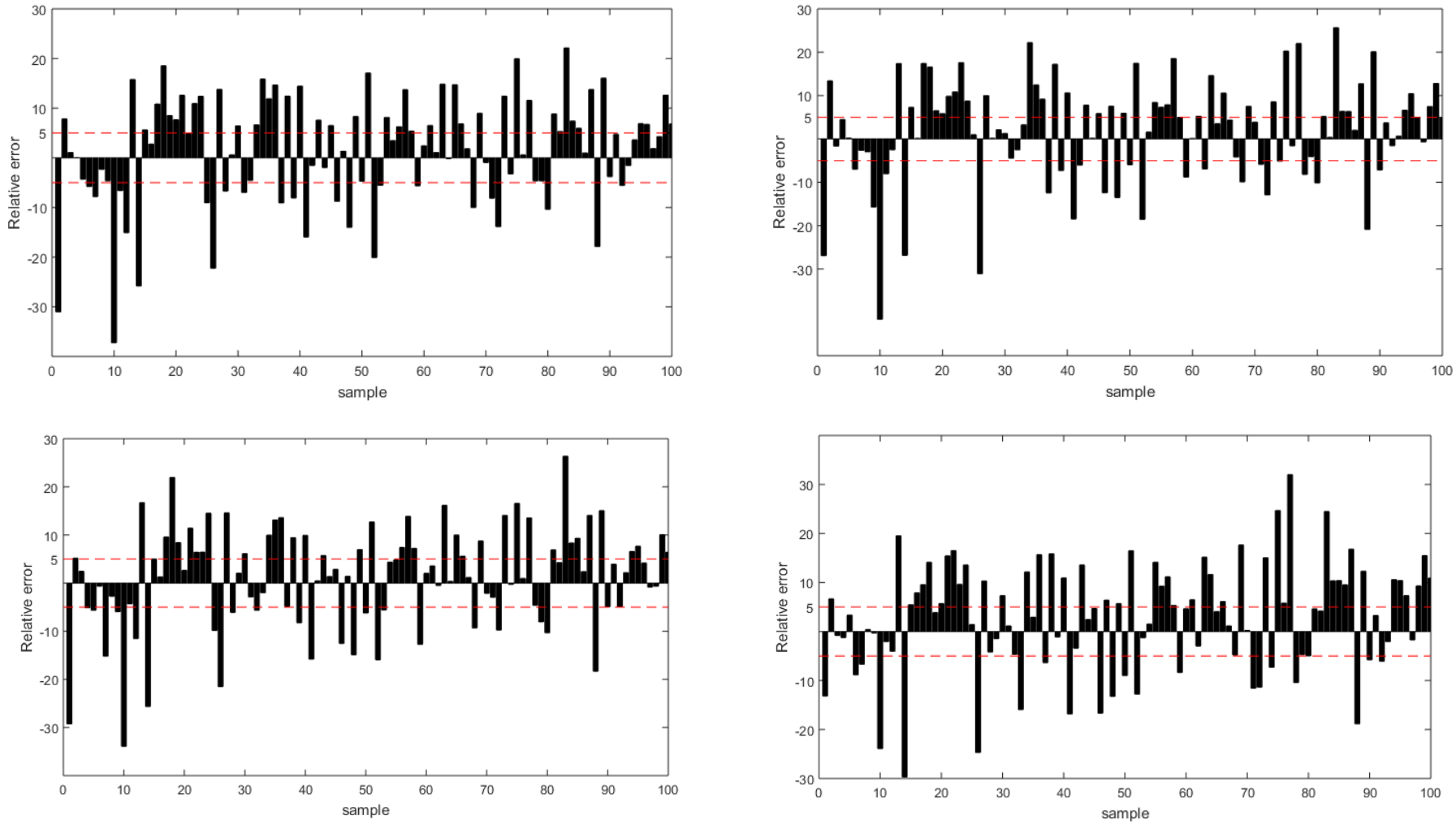


Figure 4. *Top left panel*: full model plot of relative error for each sample on the fibre content (FIB%) validation set. *Top right panel*: OPS model plot of relative error for each sample on the fibre content (FIB%) validation set. *Bottom left panel*: full model plot of relative error for each sample on the apparent sucrose in cane (PC%) validation set. *Bottom right panel*: OPS model plot of relative error for each sample on the apparent sucrose in cane (PC%) validation set.

3.2. PLS-DA results

In this study, sugarcane samples were classified based on the mean values of FIB% and PC%. In the FIB% dataset, samples that had values above the mean (13.71) were considered to as higher fibre content genotypes and assorted as belonging to the Class 1 whereas samples with values below the mean were low fibre content genotypes and were designated to Class 2. Conversely for the PC% dataset, in which the mean was 15.08. Table 7 summarizes the results obtained for the FIB% and PC% datasets.

Table 7. Performance parameters of the PLS-DA models for FIB% and PC%.

	FIB%		PC%	
	Class 1	Class 2	Class 1	Class 2
h	7		9	
Sensitivity (Cal)	0.783	0.755	0.866	0.872
Sensitivity (Pred)	0.758	0.853	1.00	0.630
Specificity (Cal)	0.755	0.783	0.872	0.866
Specificity (Pred)	0.853	0.758	0.631	1.00
Error (Cal)	0.231	0.231	0.131	0.131
Error (Pred)	0.195	0.195	0.185	0.185

h: number of latent variables; Cal: calibration set; Pred: prediction set; FIB%: fibre content; PC%: apparent sugar content.

In the best scenario, one might expect to obtain a one value for the sensitivity and specificity parameters and a null value for classification error. The classification error of FIB% was slightly higher than of PC%, 0.195, and 0.185, respectively. For PC%, the classification model assorted all samples belonging to the Class 1 correctly, therefore attained a sensitivity prediction value of 1. For FIB%, the results presented less predictability, but with a reasonable value of specificity of prediction (0.853). Therefore, the PLS-DA model developed would prevent advancing low fibre genotypes to further evaluation, but it might discard high fibre content genotypes, as a false negative result. Figure 5 shows the predicted classes of NIR-based PLS-DA models for FIB% and PC%. The condition to assort samples to one of the classes was the threshold (red dashed horizontal line). Red circles represent the samples that belong to Class 1 and black flipped squares the samples that belong to Class 2. In an ideal situation, all red circles and black flipped squares would be each on only one side of the threshold line.

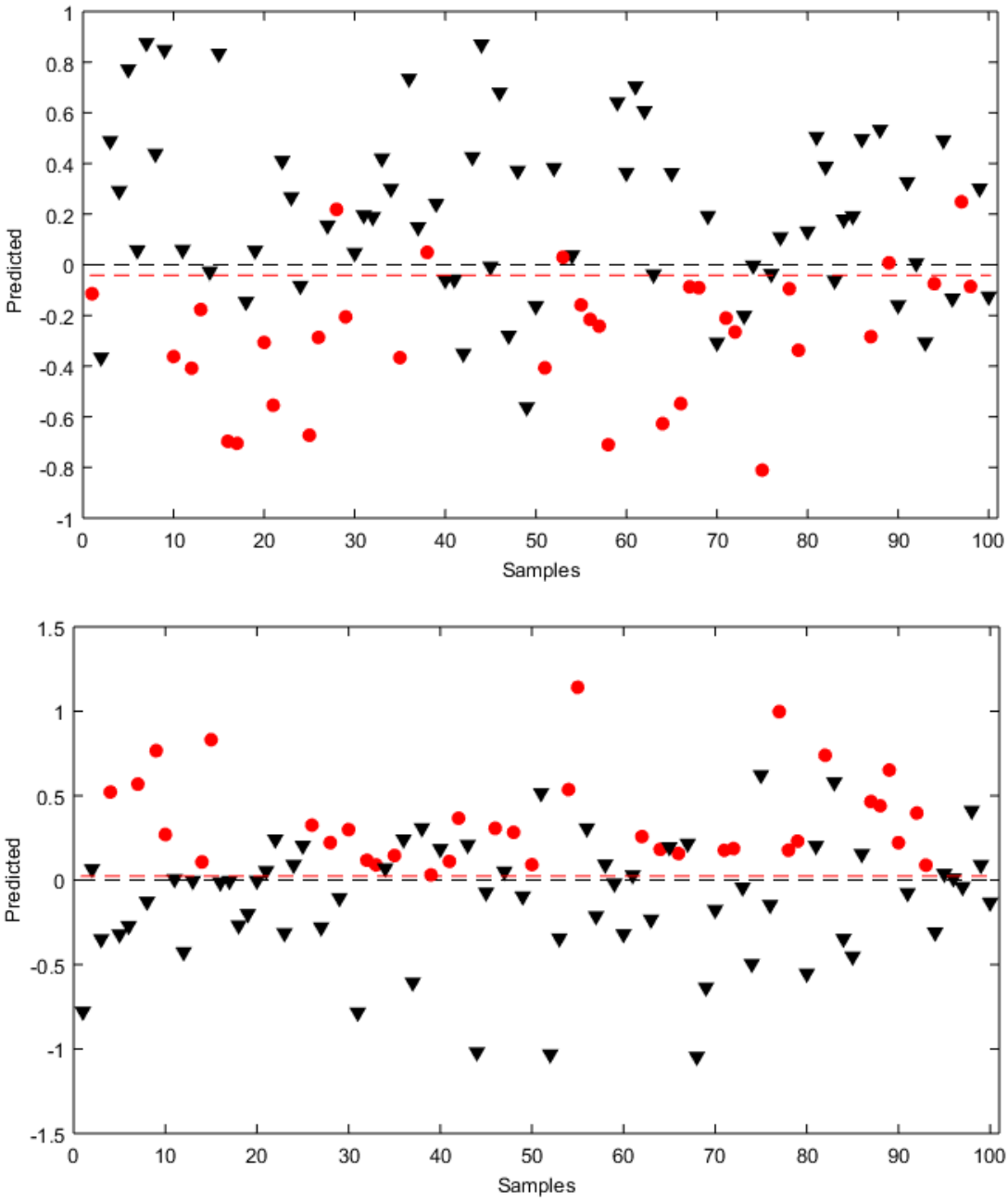


Figure 5. Predicted classes of NIR-based PLS-DA models for FIB% and PC%. *Top*: classification for fibre content (FIB%) on the validation set. *Bottom:* classification for apparent sucrose content (PC%) on the validation set.

The decision of whether it is better to correctly assign samples to a class or not is a case-by-case task. In this study for the PC% dataset, the PLS-DA model provided good discrimination ability when applied in the prediction set, as it correctly identified all genotypes possessing high apparent sucrose content.

3.3. Comparison between PLS and PLS-DA models

After splitting the data into calibration and validation sets using the Kennard-Stone algorithm, we fitted a PLS and a PLS-DA regression on samples of the calibration

set and then tested the reliability of the developed models on unknown samples of the validation set. Clones were ranked based on the true values regarding their genetic merit, determined based on the REML/BLUP approach using data of the standard chemistry analyses obtained from sugarcane samples allocated in plots of an augmented block design installed in the field. In this section, we aim to determine how well the NIR-based PLS and PLS-DA models would rank the sugarcane clones compared to the reference method currently adopted. Table 8 shows the ranking order made by the PLS model for the traits FIB% and PC% of the 30 best sugarcane clones from each corresponding validation set.

Table 8. The ranking order of the 30 best clones according to the BLUP procedure and based on the NIR-PLS approach.

	FIB%			PC%	
Clone	BLUP	NIR-PLS	Clone	BLUP	NIR-PLS
5	1º	68º	**134**	1º	19º
288	2º	6º	**116**	2º	55º
43	3º	14º	**130**	3º	20º
72	4º	54º	**69**	4º	12º
278	5º	11º	**82**	5º	84º
206	6º	81º	**14**	6º	25º
336	7º	74º	**144**	7º	11º
88	8º	21º	**96**	8º	66º
121	9º	49º	**320**	9º	23º
158	10º	44º	**314**	10º	60º
241	11º	36º	**337**	11º	72º
145	12º	16º	**84**	12º	36º
355	13º	29º	**71**	13º	24º
27	14º	51º	**293**	14º	39º
308	15º	59º	**27**	15º	14º
99	16º	82º	**295**	16º	54º
14	17º	43º	**159**	17º	5º
200	18º	2º	**46**	18º	31º
360	19º	80º	**181**	19º	8º
293	20º	39º	**229**	20º	56º
345	21º	10º	**28**	21º	75º
294	22º	8º	**200**	22º	28º
297	23º	62º	**185**	23º	35º
321	24º	18º	**321**	24º	70º
286	25º	64º	**106**	25º	37º
148	26º	24º	**13**	26º	63º
144	27º	67º	**286**	27º	81º
69	28º	20º	**296**	28º	52º
291	29º	55º	**35**	29º	65º
75	30º	40º	**287**	30º	34º

FIB%: fibre content; PC%: apparent sugar content; BLUP: best linear unbiased predictor; NIR-PLS: near infra-red partial least square model

On each validation set there were 100 samples representing the sugarcane clones, wherein we supposed the selection of the 30 best clones. Comparing the ranking lists of the two traits, we can conclude that among the 30 best clones ranked according to genotpic values, 12 clones were also among the top 30 ranked by the FIB% PLS model. On the other hand, for the PC% trait, there was a coincidence of 11 clones selected by the PLS model and BLUP procedure.

Overall, moderate accuracy values were obtained for the PLS models built for both traits, with slightly better results for the model built for predicting FIB% concentrations. A possible explanation for these results might be ascribed to the drying method we adopted. Pelletier et al. (2010) investigated the effects of different drying procedures on the forage samples chemical composition and found results indicating that the oven drying-method heavily affected the concentration of non-structural carbohydrates (e.g. sucrose). In addition, the moderate accuracies obtained for both models may be attributed to a non-linear relationship between the NIR and the analytes concentration or not enough sample variability included in the calibration set (Brereton et al., 2018). On the other hand, considering the PLS-DA models, the PC% based models outperformed the FIB% models. However, albeit the FIB% model had a moderate sensitivity value, it could deliver satisfactory results when applied in a large population. Therefore, an important advantage of this approach is the simultaneous determination of two important parameters for the industrial processing of sugarcane with the same NIR spectra.

The PMGCA is currently taking to the field approximately 30,000 clones to be assessed in the first trial phase (T1), 200 clones in the T2 phase and 30 to 60 clones in the final trial phases. The great bottleneck is the T1 phase, in which clones that will form the next subpopulations are selected. In the T1 phase no statistical design with replicates is used due to the lack of area and biological planting material. Therefore, selection is done solely based on a visual rating. The array of options after T1 is reduced as no additional genotypes are included. Additionally, a significant number of clones need to be evaluated (Brasileiro et al., 2016). In this view, the PLS-DA classification model for PC% could be an adequate option in screening sugarcane clones at the T1 trial phase of the PMGCA.

4. Conclusion

In this study we attempted to model two sugarcane feedstock quality traits using NIR spectroscopy. The PLS models developed had moderate accuracies. The correlation coefficients of prediction obtained were: 0.732 for FIB% and 0.665 for PC%. PLS models' ability to rank sugarcane clones based on their genetic merit was low, as it correctly identified only 40% for FIB% and 36% for PC% of the best genotypes. The PLS-DA models built to classify clones based on PC% showed the ideal value of 1 for sensitivity, whereas models based on FIB% showed a moderate value of 0.758. The results found in this study suggest that PLS-DA classification models using NIR spectroscopy can be a feasible tool for selecting sugarcane clones at the initial pahses of the PMGCA.

The central idea of NIR spectroscopy is that once a model for predicting a property of interest is developed it may work as a surrogate of the current standard analytical method adopted. The investigation we conducted still requires the sugarcane samples to be physically prepared, as the stalks had to be shredded, over-night dried and subsequently ground. Nevertheless, the process is less labour-intensive as it would eliminate many steps of the standard analyses. For instance, juice extraction and weighting of the fibre cake with the hydraulic press and the elimination of lead acetate for juice clarification. In addition, the models developed in this study can concurrently screen two important biomass composition traits using the same spectra. In conclusion, by adopting NIR-based models juice clarification would not be necessary, thus providing a safer environment concerning worker's health and not generating any residual.

5. References

Adami, M., Friedrich, B., Rudorff, T., Freitas, R. M., Aguiar, D. A., Sugawara, L. M., & Mello, M. P. (2012). Remote Sensing Time Series to Evaluate Direct Land Use. *Sustainability*, *4*, 574–585. https://doi.org/10.3390/su4040574

Ali, S. S., Nugent, B., Mullins, E., & Doohan, F. M. (2016). Fungal - mediated consolidated bioprocessing : the potential of Fusarium oxysporum for the lignocellulosic ethanol industry. *AMB Express*, *6*, 1–13. https://doi.org/10.1186/s13568-016-0185-0

Allwright, M. R., & Taylor, G. (2016). Molecular Breeding for Improved Second Generation Bioenergy Crops. *Trends in Plant Science*, *1*(January), 43–54. https://doi.org/10.1016/j.tplants.2015.10.002

Arruda, P. (2011). Perspective of the Sugarcane Industry in Brazil. *Tropical Plant Biology*, *8*(March), 3–8. https://doi.org/10.1007/s12042-011-9074-5

Assis, C., Ramos, R. S., Silva, L. A., Kist, V., Barbosa, M. H. P., & Teófilo, R. F. (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*, *71*(8), 2001–2012. https://doi.org/10.1177/0003702817704147

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790. https://doi.org/10.1039/c3ay40582f

Barbosa, Márcio Henrique Pereira, Resende, M. D. V. De, Peternelli, L. A., Bressiani, J. A., Silveira, L. C. I. Da, Silva, F. L., & Rodrigues, F. I. C. (2004). Use of REML/BLUP for the selection of sugarcane families specialized in biomass production. *Crop Breeding and Applied Biotechnology*, *4*(12), 218–226.

Barbosa, Marcio Henrique Pereira, Resende, M. D. V., Dias, L. A. dos S., Barbosa, G. V. de S., Oliveira, R. A., Peternelli, L. A., & Daros, E. (2012). Genetic improvement of sugar cane for bioenergy : the Brazilian experience in network research with RIDESA. *Crop Breeding and Applied Biotechnology*, *S2*, 87–98.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Bezerra, T., L., & Ragauskas, A. J. (2016). Review A review of sugarcane bagasse for second-generation bioethanol and biopower production. *Biofuels,Bioproducts & Biorefinering*, 1–14. https://doi.org/10.1002/bbb

Blanco, M., & Villarroya, I. (2002). NIR spectroscopy: A rapid-response analytical tool. *TrAC - Trends in Analytical Chemistry*, *21*(4), 240–250. https://doi.org/10.1016/S0165-9936(02)00404-1

Bordonal, R. D. O., Luís, J., Carvalho, N., Lal, R., & Figueiredo, E. B. De. (2018). Sustainability of sugarcane production in Brazil . A review. *Agronomy for Sustainable Development*, *38*(13), 1–23.

Brasileiro, B. P., Mendes, T. O. de P., Peternelli, L. A., da Silveira, L. C. I., de Resende, M. D. V., & Barbosa, M. H. P. (2016). Simulated individual best linear unbiased

prediction versus mass selection in sugarcane families. *Crop Science*, *56*(2), 570–575. https://doi.org/10.2135/cropsci2015.03.0199

Brereton, R. G. (2000). Introduction to multivariate calibration in analytical chemistry. *Analyst*, *125*(11), 2125–2154. https://doi.org/10.1039/b003805i

Brereton, Richard G. (2014). A short history of chemometrics: A personal view. *Journal of Chemometrics*, *28*(10), 749–760. https://doi.org/10.1002/cem.2633

Brereton, Richard G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., … Tauler, R. (2018). Chemometrics in analytical chemistry—part II: modeling, validation, and applications. *Analytical and Bioanalytical Chemistry*, *410*(26), 6691–6704. https://doi.org/10.1007/s00216-018-1283-4

Brereton, Richard G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics*, *28*(4), 213–225. https://doi.org/10.1002/cem.2609

Byrt, C. S., Grof, C. P. L., Furbank, R. T., & Furbank, R. T. (2011). C 4 Plants as Biofuel Feedstocks : Optimising Biomass Production and Feedstock Quality from a Lignocellulosic Perspective. *Journal of Integrative Plant Biology*, *53*(2), 120–135. https://doi.org/10.1111/j.1744-7909.2010.01023.x

Carpio, L. G. T., & De Souza, F. S. (2019). Competition between Second-Generation Ethanol and Bioelectricity using the Residual Biomass of Sugarcane: Effects of uncertainty on the production mix. *Molecules*. https://doi.org/10.3390/molecules24020369

Carvalho-Netto, O. V, Bressiani, J. A., Soriano, H. L., Fiori, C. S., Santos, J. M., Barbosa, G. V., … Leal, M. (2014). The potential of the energy cane as the main biomass crop for the cellulosic industry. *Chemical and Biological Technologies in Agriculture*, *1*(1), 20. https://doi.org/10.1186/s40538-014-0020-2

Cheavegatti-Gianotto, A., de Abreu, H. M. C., Arruda, P., Bespalhok Filho, J. C., Burnquist, W. L., Creste, S., … César Ulian, E. (2011). Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. *Tropical Plant Biology*, *4*(1), 62–89. https://doi.org/10.1007/s12042-011-9068-3

Claudio Inácio da Silveira, L., Brasileiro, B. P., Kist, V., Weber, H., Daros, E., & Peternelli, L. A. (2015). Selection strategy in families of energy cane based on biomass production and quality traits. *Euphytica*. https://doi.org/10.1007/s10681-015-1364-9

Coelho, S. T., Goldemberg, J., Lucon, O., & Guardabassi, P. (2006). Brazilian sugarcane ethanol: lessons learned. *Energy for Sustainable Development*, *10*(2), 26–39. https://doi.org/10.1016/S0973-0826(08)60529-3

CONAB (2018) Compahia Nacional de Abastecimento. (2018). Cana-de- açúcar. *Acompanhamento Da Safra Brasileira de Cana-de-Açúcar. Terceiro Levantamento - Safra 2018/9, Brasilia*, p. 71.

Consecana. (2006). *Manual de instruções (5th ed.) Piracicaba,São Paulo: Conselho doa Produtores de Cana-de-Açúcar, Açúcar e Álcool do Estado de São Paulo.*

Das, L., Liu, E., Saeed, A., Williams, D. W., Hu, H., Li, C., … Shi, J. (2017). Bioresource

Technology Industrial hemp as a potential bioenergy crop in comparison with kenaf , switchgrass and biomass sorghum. *Bioresource Technology*, *244*(June), 641–649. https://doi.org/10.1016/j.biortech.2017.08.008

de Souza, A. P., Leite, D. C. C., Pattathil, S., Hahn, M. G., & Buckeridge, M. S. (2013). Composition and Structure of Sugarcane Cell Wall Polysaccharides: Implications for Second-Generation Bioethanol Production. *Bioenergy Research*, *6*(2), 564–579. https://doi.org/10.1007/s12155-012-9268-1

Dias, M. O. S., Cunha, M. P., Jesus, C. D. F., Rocha, G. J. M., Pradella, J. G. C., Rossell, C. E. V., … Bonomi, A. (2011). Second generation ethanol in Brazil: Can it compete with electricity production? *Bioresource Technology*, *102*(19), 8964–8971. https://doi.org/10.1016/j.biortech.2011.06.098

Dias, M. O. S., Junqueira, T. L., Jesus, C. D. F., Rossell, C. E. V, Maciel, R., & Bonomi, A. (2012). Improving second generation ethanol production through optimization of fi rst generation production process from sugarcane. *Energy*, *43*(1), 246–252. https://doi.org/10.1016/j.energy.2012.04.034

Emanuel Fernando Maia de, S., Luiz Alexandre, P., & Márcio Henrique Pereira, B. (2006). Designs and model effects definitions in the initial stage of a plant breeding program Delineamentos e definições de efeitos no modelo em estágios iniciais de melhoramento vegetal. *Pesq. Agropec. Bras.*, (1), 369–375. Retrieved from /scielo.php?script=sci_arttext&pid=&lang=en

Federer, W. T. (1961). AUGMENTED DESIGNS WITH ONE-WAY ELIMINATION OF HETEROGENEITY. *Biometrics*, *17*(9), 447–473.

Fernández-coppel, I. A., Barbosa-evaristo, A., Corrêa-guimarães, A., Martín-gil, J., Navas-gracia, L. M., & Martín-ramos, P. (2018). Industrial Crops & Products Life cycle analysis of macauba palm cultivation : A promising crop for biofuel production. *Industrial Crops & Products*, *125*(August), 556–566. https://doi.org/10.1016/j.indcrop.2018.09.036

Ferreira, M. M. C., Antunes, A. M., Melgo, M. S., & Volpe, P. L. O. (1999). Quimiometria i: Calibração multivariada, um tutorial. *Quimica Nova*, *22*(5), 724–731. https://doi.org/10.1590/S0100-40421999000500016

Galvão, R. K. H., Araujo, M. C. U., José, G. E., Pontes, M. J. C., Silva, E. C., & Saldanha, T. C. B. (2005). A method for calibration and validation subset partitioning. *Talanta*, *67*(4), 736–740. https://doi.org/10.1016/j.talanta.2005.03.025

Henderson, A. C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, *31*(2), 423–447.

Henry, R. J. (2010). Evaluation of plant biomass resources available for replacement of fossil oil. *Plant Biotechnology Journal*, *8*, 288–293. https://doi.org/10.1111/j.1467-7652.2009.00482.x

Hoang, N. V, Furtado, A., Donnan, L., Keeffe, E. C., Botha, F. C., & Henry, R. J. (2017). *High-Throughput Profiling of the Fiber and Sugar Composition of Sugarcane Biomass*. 400–416. https://doi.org/10.1007/s12155-016-9801-8

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior*

Research Methods, *48*(4), 1308–1317. https://doi.org/10.3758/s13428-015-0624-x

Ildiko, E. F., & Jerome, H. F. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, *35*(2), 109–135.

Jackson, P. A. (2005). Breeding for improved sugar content in sugarcane. *Field Crops Research*, *92*, 277–290. https://doi.org/10.1016/j.fcr.2005.01.024

James, G., Witen, D., Hastie, T., & Tibshirani, R. (2007). *An Introduction to Statistical Learning with Applications in R*. https://doi.org/10.1016/j.peva.2007.06.006

Kandel, R., Yang, X., Song, J., & Wang, J. (2018). Potentials , Challenges , and Genetic and Genomic Resources for Sugarcane Biomass Improvement. *Frontiers in Plant Science*, *9*(February), 1–14. https://doi.org/10.3389/fpls.2018.00151

Kennard, R. W., & Stone, L. A. (1969). Technometrics Computer Aided Design of Experiments. *Technometrics*, *11*(1), 137–148. https://doi.org/https://doi.org/10.1080/00401706.1969.10490666

Kim, M., & Day, D. F. (2010). Composition of sugar cane , energy cane , and sweet sorghum suitable for ethanol production at Louisiana sugar mills. *Journal of Industrial Microbiology & Biotechnology*, *38*, 803–807. https://doi.org/10.1007/s10295-010-0812-8

Kjeldahl, K., & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, *24*(7–8), 558–564. https://doi.org/10.1002/cem.1346

Laurance, W. F. (2007). Switch Corn Promotes Amazon Deforestation. *Science*, *318*, 1721.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst*, *143*(15), 3526–3539. https://doi.org/10.1039/c8an00599k

Legendre, B. L., & Burner, D. M. (1995). BIOMASS PRODUCTION OF SUGARCANE CULTIVARS AND EARLY-GENERATION HYBRIDS. *Biomass and Bioenergy*, *8*(2), 55–61.

Leite, M. S. D. O., Peternelli, L. A., & Barbosa, M. H. P. (2006). Effects of plot size on the estimation of genetic parameters in sugarcane families. *Crop Breeding and Applied Biotechnology*, *6*(1), 40–46.

Lewandowski, I. Ã., & Faaij, A. P. C. (2006). Steps towards the development of a certification system for sustainable bio-energy trade. *Biomass and Bioenergy*, *30*, 83–104. https://doi.org/10.1016/j.biombioe.2005.11.003

Lopes, M. L., Paulillo, S. C. de L., Godoy, A., Cherubin, R. A., Lorenzi, M. S., Giometti, F. H. C., … de Amorim, H. V. (2016). Ethanol production in Brazil: a bridge between science and industry. *Brazilian Journal of Microbiology*, *47*, 64–76. https://doi.org/10.1016/j.bjm.2016.10.003

Macedo, I. C., Seabra, J. E. A., & Silva, E. A. R. (2008). Green house gases emissions in the production and use of ethanol from sugarcane in Brazil : The 2005 / 2006 averages and a prediction for 2020. *Biomass and Bioenergy*, *32*(7), 582–595. https://doi.org/10.1016/j.biombioe.2007.12.006

Marin, Fa. R., Martha, G. B. J., Cassman, K. G., & Grassini, P. (2016). Prospects for Increasing Sugarcane and Bioethanol Production on Existing Crop Area in Brazil. *BioScience*, *66*(4), 307–316. https://doi.org/10.1093/biosci/biw009

Matsuoka, S., Ferro, J., & Arruda, P. (2009). The Brazilian Experience of Sugarcane Ethanol Industry The Brazilian experience of sugarcane ethanol industry. *In Vitro Cellular & Developmental Biology*, *45*(January), 372–381. https://doi.org/10.1007/978-1-4419-7145-6

Matsuoka, S., Kennedy, A. J., Gustavo, E., Santos, D., Tomazela, A. L., & Rubio, L. C. S. (2014). Energy Cane : Its Concept , Development , Characteristics , and Prospects. *Advances in Botany*, *2014*, 13.

Mendu, V., Shearin, T., Campbell, J. E., Stork, J., Jae, J., Crocker, M., … DeBolt, S. (2012). Global bioenergy potential from high-lignin agricultural residue. *Proceedings of the National Academy of Sciences*, *109*(10), 4014–4019. https://doi.org/10.1073/pnas.1112757109

Moosavi, S. A., Aghaalikhani, M., Ghobadian, B., & Fayyazi, E. (2018). Okra : A potential future bioenergy crop in Iran. *Renewable and Sustainable Energy Reviews*, *93*(April), 517–524. https://doi.org/10.1016/j.rser.2018.04.057

Nawi, N. M., Rowshon, K. M., Guangnan, C., & Troy, J. (2014). Prediction of Sugarcane Quality Parameters Using Visible-shortwave Near Infrared Spectroradiometer. *Agriculture and Agricultural Science Procedia*, *2*, 136–143. https://doi.org/10.1016/j.aaspro.2014.11.020

Nepstad, D., McGrath, D., Stickler, C., Alencar, A., Azevedo, A., Swette, B., … Hess, L. (2014). Slowing Amazon deforestation through public policy and interventions in beef and soy supply chains. *Science*, *344*(6), 1118–1123. https://doi.org/10.1126/science.1248525

Pasquini, C. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, *14*(2), 198–219. https://doi.org/10.1590/S0103-50532003000200006

Pasquini, Celio. (2018). Analytica Chimica Acta Near infrared spectroscopy : A mature analytical technique with new perspectives e A review. *Analytica Chimica Acta*, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004

Pelletier, S., Tremblay, G. F., Bertrand, A., Bélanger, G., Castonguay, Y., & Michaud, R. (2010). Drying procedures affect non-structural carbohydrates and other nutritive value attributes in forage samples. *Animal Feed Science and Technology*, *157*(3–4), 139–150. https://doi.org/10.1016/j.anifeedsci.2010.02.010

Phuphaphud, A., Saengprachatanarug, K., Posom, J., Maraphum, K., & Taira, E. (2019). Prediction of the fibre content of sugarcane stalk by direct scanning using visible-shortwave near infrared spectroscopy. *Vibrational Spectroscopy*, *101*(August 2018), 71–80. https://doi.org/10.1016/j.vibspec.2019.02.005

Popp, J., Lakner, Z., Harangi-rákos, M., & Fári, M. (2014). The effect of bioenergy expansion : Food , energy , and environment. *Renewable and Sustainable Energy Reviews*, *32*, 559–578. https://doi.org/10.1016/j.rser.2014.01.056

Porto, N. de A., Roque, J. V., Wartha, C. A., Cardoso, W., Peternelli, L. A., Barbosa, M.

H. P., & Teófilo, R. F. (2019). Early prediction of sugarcane genotypes susceptible and resistant to Diatraea saccharalis using spectroscopies and classification techniques. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *218*, 69–75. https://doi.org/10.1016/j.saa.2019.03.114

Purcell, D. E., O'shea, M. G., Johnson, R. A., & Kokot, S. (2009). Near-infrared spectroscopy for the prediction of disease ratings for fiji leaf gall in sugarcane clones. *Applied Spectroscopy*, *63*(4), 450–457. https://doi.org/10.1366/000370209787944370

Ramos, R. S., Brasileiro, B. P., Kist, V., Assis, C., Gasparini, K., Silva, L. A., … Peternelli, L. A. (2017). Selection of energy cane clones. *Crop Breeding and Applied Biotechnology*, *17*, 327–333.

Resende, M. D. V. (2016). Software Selegen-REML/BLUP: A useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*, *16*(4), 330–339. https://doi.org/10.1590/1984-70332016v16n4a49

Rinnan, Å., Berg, F. Van Den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roque, J. V., Dias, L. A. S., & Teófilo, R. F. (2017). Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies. *Jorunal of the Brazilian Chemistry Society*, *28*(8), 1506–1516.

Sabatier, D., Dardenneb, P., & Thuriès, L. (2011). Near infrared reflectance calibration optimisation to predict lignocellulosic compounds in sugarcane samples with coarse particle size. *Journal of Near Infrared Spectroscopy*, *19*(3), 199–209. https://doi.org/10.1255/jnirs.929

Sang, T. (2011). Toward the Domestication of Lignocellulosic Energy Crops : Learning from Food Crop Domestication. *Journal of Integrative Plant Biology*, *53*(2), 96–104. https://doi.org/10.1111/j.1744-7909.2010.01006.x

Santchurn, D., Ramdoyal, K., Badaloo, M. G. H., & Labuschagne, M. T. (2014). ScienceDirect From sugar industry to cane industry : Evaluation and simultaneous selection of different types of high biomass canes. *Biomass and Bioenergy*, *61*, 82–92. https://doi.org/10.1016/j.biombioe.2013.11.023

Silva, A. L., Gasparini, K., Assis, C., Ramos, R., Kist, V., Barbosa, M. H. P., … Bhering, L. L. (2017). Selection strategy for indication of crosses between potential sugarcane genotypes aiming at the production of bioenergy Lidiane. *Industrial Crops & Products*, *104*, 62–67. https://doi.org/10.1016/j.indcrop.2017.04.025

Sims, R. E. H. (2004). Renewable energy : a response to climate change. *Solar Energy*, *76*, 9–17. https://doi.org/10.1016/S0038-092X(03)00101-4

Somerville, C., Somerville, C., Youngs, H., Taylor, C., Davis, S. C., & Long, S. P. (2010). Feedstock for Lignocellulocis Biofuels. *Science*, *323*(5993), 790–792. https://doi.org/10.1126/science.1189268

Sticklen, M. (2006). Plant genetic engineering to improve biomass characteristics for biofuels. *Current Opinion in Biotechnology*, *17*(3), 315–319.

https://doi.org/10.1016/j.copbio.2006.05.003

Stringer, J. K., Cox, M. C., Atkin, F. C., Wei, X., & Hogarth, D. M. (2011). Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. *Sugar Tech*, *13*(1), 36–41. https://doi.org/10.1007/s12355-011-0073-5

Surendra, K. ., Ogoshi, R., Zaleski, H. ., Hashimoto, A. ., & Khanal, S. K. (2018). High yielding tropical energy crops for bioenergy production: Effects of plant components, harvest years and locations on biomass composition. *Bioresource Technology*, *251*(3), 218–229. https://doi.org/10.1016/j.biortech.2017.12.044

Teófilo, R. F., Martins, J. P. A., & Ferreira, M. M. C. (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics*, *23*(1), 32–48. https://doi.org/10.1002/cem.1192

Valderrama, P., Braga, J. W. B., & Jesus, P. R. (2007). Variable Selection, Outlier Detection, and Figures of Merit Estimation in a Partial Least-Squares Regression Multivariate Calibration Model. A Case Study for the Determination of Quality Parameters in the. *Journal of Agricultural and Food Chemistry*, *55*, 8331–8338.

Valderrama, P., Braga, J. W. B., & Poppi, R. J. (2007). Validation of Multivariate Calibration Models in the Determination of Sugar Cane Quality Parameters by Near Infrared Spectroscopy. *Journal of Brazilian Chemistry Society*, *18*(2), 259–266.

Waclawovsky, A. J., Sato, P. M., Lembke, C. G., Moore, P. H., & Glaucia, M. (2010). Sugarcane for bioenergy production : an assessment of yield and regulation of sucrose content. *Plant Biotechnology Journal*, *8*, 263–276. https://doi.org/10.1111/j.1467-7652.2009.00491.x

Wang, L. P., Jackson, P. A., Lu, X., Fan, Y. H., Foreman, J. W., Chen, X. K., … Aitken, K. S. (2008). Evaluation of sugarcane x Saccharum spontaneum progeny for biomass composition and yield components. *Crop Science*, *48*(3), 951–961. https://doi.org/10.2135/cropsci2007.10.0555

Wang, Y., Fan, C., Hu, H., Li, Y., Sun, D., & Peng, L. (2016). Genetic modification of plant cell walls to enhance biomass yield and biofuel production in bioenergy crops. *Biotechnology Advances*, *34*(5), 997–1017. https://doi.org/10.1016/j.biotechadv.2016.06.001

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, *44*(1–2), 175–185. https://doi.org/10.1016/S0169-7439(98)00109-9

Wold, S., Sjostrom, M., Eriksson, L., & Sweden˚˚, S. (2001). PLS-regression, a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 2001–2109. https://doi.org/10.1016/S0169-7439(01)00155-1

Zhang, L., Li, G., Sun, M., Li, H., Wang, Z., Li, Y., & Lin, L. (2017). Kennard-Stone combined with least square support vector machine method for noncontact discriminating human blood species. *Infrared Physics and Technology*, *86*, 116–119. https://doi.org/10.1016/j.infrared.2017.08.020

CHAPTER 2

COMBINING PHENOMIC AND GENOMIC INFORMATION FOR SUGARCANE BREEDING

1. Introduction

International commitment policies such as the Kyoto Protocol and the Paris agreement, which aimed to mitigate climate change effects, have renewed the interest for biomass-based energy production (Mendu et al., 2012; Hanegraaf et al., 1998). To this end, sugarcane (*Saccharum spp*.) is highlighted as a candidate plant species. Sugarcane is an important industrial crop cultivated in several countries, notably in Brazil. The Brazilian sugarcane industry has been supported - although with no desirable consistency - by the Brazilian government since 1975 with the National Alcohol Program (Proálcool). The Proálcool program fostered the sugarcane industry providing financial and incentive policies. These actions have helped strengthen the sector and led to scientific research and technological innovation that established the concept of biorefinery and cogeneration systems, in which sugarcane feedstock is efficiently converted into bioethanol or burned to produce steam and generate electricity (Matsuoka et al., 2009). Within the scope of agriculture practices and biotechnology, breeding programs play a central role in the economic sustainability of the sugarcane industry. Barbosa (2012) estimates that 50 % of gains in productivity in sugarcane fields are the result of the substitution of less suitable cultivars. Therefore, the continuous release of more adapted, high yield, and pest-resistant cultivars has been a major factor of the success of Brazilian mills and distilleries (Lopes et al., 2016).

Sugarcane is a perennial crop, and its cultivation is made through clonal propagation. Extensive field trials over many years and breeding cycles are required before the release of a new cultivar. Family and individual selection based on phenotypic data are the main strategies adopted by sugarcane breeders. The whole procedure can take up to 15 years. Sugarcane breeding commences with the selection and crossing of suitable genitors and the subsequent assessment in the field of the seedlings generated. The first evaluation trial is named T1 phase (Brasileiro et al., 2016). At the T1 phase mass selection is employed, in which clones are screened only by a visual appraisal. However, it is well known that this approach is not satisfactory, as the environment and other external factors may bring bias to the evaluation, especially when low heritability traits are considered (Langridge et al., 2011). Consequently, bad genotypes may be favoured by being located in an ideal spot and advanced. Conversely, good genotypes may be discarded. Therefore, even an experienced breeder can be inclined to making erroneous decisions.

Phenotyping is a laborious, costly, and time-consuming task. Additionally, phenotyping is prone to human error, and therefore, the genetic potential of populations may not be fully exploited. Indeed, ever since the Danish plant physiologist and botanist Wilhelm Johannesen coined the term phenotype, field-based collection of data, commonly referred to phenotyping, has been a cornerstone for plant breeders (Berry, 2014). In short, at the sugarcane genetics breeding program of the Universidade Federal de Viçosa (PMGCA), two major problems need to be circumvented. Firstly, the lack of seed material in the T1 phase, which is to say stalk cuttings (setts) used to propagate sugarcane clones. That limitation prevents the conduction of more refined experimental designs with replications that could ultimately improve the inference over a breeding population. Moreover, better tools to efficiently screen sugarcane clones at early phases of selection are necessary. Motivated by this, we set out to find solutions and strategies able to tackle these constraints.

The emergence of new technologies is reshaping the way data is collected. We are already living in the age of big data as cameras, sensors, and other devices are continuously collecting and storing digital information. This phenomenon is mainly happening at industrial facilities, tech companies, and government institutions. Nevertheless, agriculture research is not an exception. New high-throughput platforms both at the genome and phenome level are currently undergoing advancements and show a downward trend in costs and time required to be performed. Therefore, the adoption of high-throughput phenotyping and next-generation sequencing technologies is heralded to be commonplace at breeding and research stations in the years to come.

Phenomics can be defined as the efficient and large-scale collection of phenotypes, which is to say high-throughput phenotyping (HTP; Deshmukh et al., 2014). Phenomics is an incipient, though a growing area of interest among plant breeders (Araus et al., 2014). If we discard the occurrence of mutations and suppose epigenetic changes or other related mechanism are absent, the genome of a plant is immutable. On the other hand, the phenome is ever-changing and thereby cannot be completely characterized (Houle et al., 2010). Nevertheless, phenomic data can have great utility at breeding programs. For instance, HTP platforms-derived data of traits like drought tolerance or disease resistance can be linked to gene expression in longitudinal studies and aid in better understand the underlying phenomena of genetic variation affecting complex traits (Montes et al., 2007; Singh et al., 2016). Moreover, HTP technologies can replace standard ineffective or

deficient phenotyping protocols, thus saving much time and resources. NIR spectroscopy is an example that has already been successfully applied to screen biological samples composition and for breeding purposes (Purcell et al., 2009; Roggo et al., 2004; Roque et al., 2017).

Concurrent with HTP platforms, the advent of DNA-based molecular markers coupled with improved instruments and laboratory techniques have allowed the emergence of next-generation sequencing (NGS; Goodwin et al., 2016; Schlotterer, 2004). NGS technologies can deliver DNA-level information at an ever more cost-effective and high-throughput manner. In plant research, genomic enabled prediction was initially conceived using molecular markers to select genotypes based on the identification and estimation of quantitative trait loci (QTL) associated with phenotypes, e.g., disease resistance. This approach is referred to as marker-assisted selection (MAS). However, MAS is unable to fully explain the genetic variation of complex traits. This limitation lays on the quantitative nature of several agronomical important traits, as they are assumed to be governed by many genes, each with small effects. Thus, it is expected that both major and minor QTLs play a role in the determination of complex traits (Crossa et al., 2017).

Contrary to MAS, genomic selection utilizes all available molecular markers. The use of all available molecular markers in a joint manner to estimate the genetic merit of individuals was foreseeing by Meuwissen et al. (2001) even before dense genome-wide molecular markers were available. In the literature, the approach is termed interchangeably as whole-genome, genome-wide, and genomic selection. The idea of genomic selection (GS) is to fit a regression model using the entire set of available molecular markers and phenotypic values obtained from a training population. After, the developed model is tested on a test population panel. The model developed might allow the prediction of plant or animals genetic merit of genotyped but nonphenotyped populations (De Los Campos et al., 2013).

The genetic value estimated using the GS approach is referred to as genomic estimated breeding value (GEBV). In most statistical strategies devised to handle genomic data only the additive components of the genetic variation are assumed (Sant' Anna et al., 2019). Thus, the phenotypic response of each individual takes the form of a standard linear model for n genotypes ($i = 1\ldots, n$) and p molecular markers ($j = 1\ldots, p$), and can be written as:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i ,$$

where y_i is the phenotypic observation for the *i*th genotype, x_{ij} is the molecular marker covariates, β_j is the effect of the *j*th marker and ε_i is the residual (attributed to everything that is not due to the genetic term) for the *i*th genotype. Once marker effects are estimated, the GEBV is computed as follows (Xavier et al., 2016):

$$GEBV = X\hat{\beta} ,$$

where $\hat{\beta}$ is the vector containing the molecular marker effects and *X* is the matrix of the genotyped breeding population panels (Gouy et al., 2013). Genomic selection is aimed to reduce generation interval as genitors could be crossed, and their progeny have their DNA collected from seeds or seedlings and genotyped. Empirical studies suggest that the genetic gain per cycle is slightly superior when compared to conventional breeding (Heffner et al., 2010). However, the gains per unit of time are substantially increased. Therefore, the adoption of genomic selection is expected to augment selection efficiency. Consequently, efforts and resources can be allocated to other activities. The reduction of costs is predicted to be higher in perennial crops (Bernardo, 2008; Resende et al., 2012). GS is ever more winning its place at sugarcane breeding programs. Indeed, GS is already standard practice in private companies (Mammadov et al., 2014). However, only a limited set of species have been receiving attention, and further investigation on polyploidy crops such as sugarcane is needed (Allwright et al., 2016; Hoang et al., 2015)

The startling amount of data being generated is outpacing our ability to explore it. Owing to this fact, a question arises: how can breeders properly harness the information and effectively apply it in their practical work. To this date, the integration of omics data is being proposed by the exploitation of omics databases and attempting to develop gene regulatory networks or as a decision support system (Deshmukh et al., 2014) or by recording covariates using HTP platforms and molecular markers for the creation of a selection index (Mackay et al., 2015). However, in this study, we propose to investigate whether a NIR spectra dataset - a HTP platform - and a genome-wide molecular marker array can be directly integrated into a multi-omics analyses. The rationale behind our thesis is that by adding more covariates to the single nucleotide polymorphism (SNP)

array, it may provide a higher prediction accuracy of key complex traits. Jannink et al. (2010) have already highlighted the similarity of the NIR spectroscopy and GS approaches, as they inherently share the same purposes and statistical analysis challenges. NIR spectroscopy aims to replace demanding and expensive laboratory protocols by developing prediction models using multivariate statistics. Likewise, GS is intended to utilize multivariate statistic methods to build prediction models that associate tiresome to record plant phenotypes with easy to measure variables. Moreover, in both procedures, the input matrices are high dimensional datasets that present high levels of multicollinearity and therefore impair the application of traditional fitting procedures such as ordinary least squares (OLS; Crossa et al., 2010). The existing statistical methods able to cope with these challenges are used across fields (Boulesteix et al., 2007; Gouy et al., 2013). For instance, Silveira et al. (2016) applied partial least squares (PLS) regression in a genomic selection study using SNPs to predict pH in pork meat after slaughter. To the best of our knowledge, there is no published study which proposes the concatenation of NIR spectra and molecular markers data sets. In this chapter we aimed to determine if the accuracy of genomic selection for two traits in a sugarcane population genotyped using SNPs markers is improved by combining the NIR spectra matrix to a SNP genotyping matrix into a single regression analysis.

2. Material and methods

The sugarcane clones used in this study were derived from a breeding population of the Universidade Federal de Viçosa Sugarcane Genetic Breeding Program (PMGCA) and are identical to the ones used in Chapter 2. However, we had missing samples, and miss correspondence between individuals that had both their NIR spectra collected and SNP genotyped. Thus, in this study, the population consisted of three-hundred and eighty-five sugarcane clones. Likewise, the experimental design, sample preparation, and NIR spectra acquisition are the same as Chapter 1, in sections 2.2, 2.3, and 2.4. Sugarcane clones were phenotyped for fibre (FIB%) and apparent sucrose content (PC%). The phenotypic data was analysed by mixed model equation means using the software Selegen-REML/BLUP (Resende, 2016) and the estimated genetic values were used as dependent variables of the y_{nx1} vector. In Figure 1**,** it is displayed the plots of the genotypic value estimates for the two evaluated traits and measured phenotypic values. It

is evident that the correlation is high, and therefore, we assumed it would be equivalent to use either value in the fitting procedures.

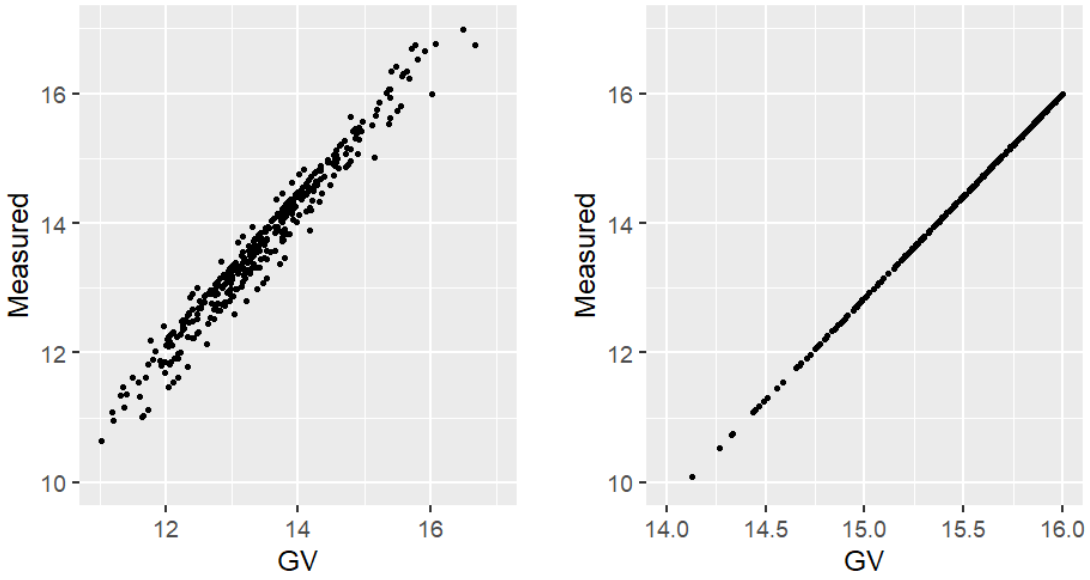


Figure 1. *Left:* Plot of genotypic value (GV) estimates versus measured values for FIB%. *Right*: Plot of genotypic values estimates versus and measured values for PC%.

2.1. DNA isolation, sequencing, and genotyping analyses

Sugarcane DNA samples were isolated using DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany) and sent to RAPiD Genomics (Gainesville, Florida, USA) for the construction of probes, sequencing, and identification of molecular markers. Samples were genotyped using single-dose SNP markers based on the Capture-seq technology (https://www.rapid-genomics.com). Raw sequence reads were mapped, called, and filtered. Reads were anchored to a monoploid reference genome of sugarcane (*Saccharum spp.*; Garsmeur et al., 2018) using BWA-MEM algorithm of BWA version 0.7.17 (Li et al., 2010), and a flag identifying the respective sugarcane genotype was added to each mapping file. Next, the mapping files were processed using SortSam, MarkDuplicates, and BuildBamIndex tools of Picard version 2.18.27 (https://github.com/broadinstitute/picard/). Variants were called using FreeBayes version 1.2.0 (https://github.com/ekg/freebayes) with a minimum mapping quality of 20 (probability of miscalling), minimum base quality of 20 (SNPs with missing data higher than 20% were eliminated), and minimum coverage (how many times a fragment was sequenced) of 20 reads at every position in the reference genome. Thereafter, the SNP marker matrix was coded counting the occurrence of the reference allele **A**. Thus, considering the

genotypes **AA**, **Aa**, and **aa**, the matrix entries would be **2** (homozygosity for the reference allele), **1** (heterozygosity with one reference and one alternative allele), and **0** (homozygosity for the alternative allele), respectively. Further, markers with minor allele frequency (MAF) higher than 10% were eliminated. Lastly, the genotyping matrix was submitted to imputation of missing genotypes from a binomial distribution density function using values of the non-missing genotypes. The final matrix used in the analyses had 385 rows and 124340 columns.

2.2. Regression models

2.2.1. Bayesian lasso (BLASSO)

The BLASSO is a Bayesian counterpart of the least absolute shrinkage and selection operator (LASSO; Park et al., 2008). The LASSO is classified as a penalized regression method, which is to say, the fitting procedure involves minimizing the residual sum of squares subjected to a penalization (Tibshirani, 1996). Likewise, the regression coefficient estimates of the BLASSO are assumed to have random effects and a double-exponential prior distribution (Pérez et al., 2014). The solution is obtained by solving an optimization problem that takes the following form:

$$\hat{\beta} = \text{argmin}\left\{\sum_{i}^{n}\left(y_i - \sum_{j}^{p} x_{ij}\beta_j \right)^2 + \lambda\sum_{j}^{p}\left|\beta_j\right|\right\}$$

where y_i is the phenotypic response of the *ith* individual, x_{ij} is the genotype of the *ith* individual at the *jth* marker (x_{ij} = 2 for AA; x_{ij} = 1 for Aa; x_{ij} = 0 for aa), β_j is the allelic substitution effect for the *jth* marker and λ is the penalization parameter that controls the trade-off between model complexity and goodness of fit (Usai et al., 2009). The Gibbs sampler is an algorithm applied to obtain the regression coefficients estimates (that are stored and averaged at the end of the process).

2.2.2. Partial least square (PLS) regression

The most usual method to estimate parameters in linear regression is OLS, wherein the unknown coefficient estimates are obtained from a general linear model, solving the normal equation system:

$$X'X\hat{\beta} = X'Y$$
$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

However, in situations wherein the assumptions used for OLS do not hold, e.g. rank-deficiency problems ascribed to the existing multicollinearity and high-dimensionality, other strategies are employed. In the PLS regression approach, a generalized inverse is applied when finding the solution of the system. In addition, the matrix X is factorized as following:

$$Y_{nx1} = X_{nxm}\beta_{mx1}$$
$$Y_{nx1} = \left(U_{nxh}S_{hxh}V'_{hxm}\right)\beta_{mx1}$$

After factorization, the matrix is truncated by choosing the number of latent variables. Then, the Moore-Penrose pseudoinverse is applied, and regression coefficients can be computed:

$$\hat{X}^{+}_{mxn} = V_{mxh}\left(S_{hxh}\right)^{-1}U'_{hxn}$$
$$\hat{\beta}_{mx1} = \hat{X}^{+}_{mxn}Y_{nx1}\ ,$$

where $\hat{X}^{+}$ is the generalized inverse of X after factorization and h is the number of latent variables (Brereton et al., 2018). The PLS method identifies the principal components (latent variables) that best describe the data in terms of variance, and it does so by constructing linear combinations of all predictors. Furthermore, unlike principal component regression (PCR), the fitting procedure of PLS involves finding the latent variables that maximize the covariance between the predictor and response variables while minimizing the error (James et al., 2007; Roque et al., 2017).

2.3. Accuracy of predictions and matrix concatenation

The prediction accuracies of models were tested by partitioning the data into two parts using the Kennard-stone (KS) algorithm. KS aims to select representative subsamples sets. The algorithm maximizes the Euclidean distances $Dx(p,q)$ between samples in the instrument output matrix's vectors:

$$Dx(p,q) = \sqrt{\sum_{j=1}^{J}[\mathrm{x}_p(\mathrm{j}) - \mathrm{x}_q(\mathrm{j})]^2}\ ,\ p,q \in [1,\mathrm{N}]$$

where N is the total number of samples and $x_p(j)$, and $x_q(j)$ are the absorbance values at the *jth* wavelength for the pair of samples p and q (Galvão et al., 2005). The KS algorithm was applied in the genotyping matrix alone and after, using the same selected samples in the NIR matrix. Then, the dataset was split into training and validation sets. The training set contained 80% of the samples (308 clones), and the validation set contained the remainder 20% (77 clones). Hence, only one partitioning was done. The final matrices had 385 rows and 127340 (124340 + 3112) columns. The models were fitted using the data of the training observations and tested on unknown samples of the validation set. The results were compared by computing the Pearson correlation coefficient between true breeding values (TBV) values and predicted (GEBV) by genomic selection models.

2.4. Statistical analysis

The models using the genomic and NIR datasets were fitted using the following linear model:

$$y = \mu + Xb + Wa + \varepsilon$$

where y is the phenotype vector (FIB% and PC% genotypic values), μ is the intercept, X is a design matrix of spectra wavelengths, b is the corresponding vector of effects, W is the matrix with marker genotypes, a is the corresponding vector of marker effects and ε is the vector of residuals. We considered a and b vectors of random effects and assigned double-exponential priors. The statistical analyses were carried out in R software using in-house developed scripts and following the guidelines of the packages BGLR (Pérez; et al., 2014). The BLASSO algorithm was run for 50,000 cycles with the first 25,000 being discarded (burn-in) and with a thin interval of 10. In addition, PLS regression was performed in R software using the pls package.

3. Results

We repeated the analyses of the NIR dataset using three hundred and eighty-five observations and assessed the performance of PLS and BLASSO regression methods (Table 1). The PLS model was fitted using seven latent variables for FIB% and 11 latent variables for PC%. The pretreatments that yielded the best results were Savitzky-Golay smoothing (window: 5; polynomial order: 2), 1º derivative, multiplicative scatter correction and mean centre for FIB% (Engel et al., 2013). For PC% the best combination of pretreatments were Savitzky-Golay smoothing (window: 5; polynomial order: 2) and

mean centre. In this analyses the Gibbs sampler was setting as follow: 25,000 iterations, a burn-in of 10,000 and thin interval of 10. The results found in Table 1 suggest that the performance of BLASSO and PLS are quite similar. Therefore, we assumed the BLASSO would fit the analyses for the concatenated matrix.

Table 1. Performance of BLASSO and PLS regression in the NIR dataset.

	FIB%		PC%	
	R_p	RMSEP	R_p	RMSEP
PLS	0.570	0.950	0.600	1.560
BLASSO	0.552	0.924	0.550	1.660

FIB%: fibre content; PC%: apparent sucrose content; PLS: partial least squares; BLASSO: Bayesian least absolute shrinkage selection operator; R_p: correlation coefficient of prediction; RMSEP: root mean squared error of prediction.

3.1 Genomic selection

In this study, genomic selection was applied to identify SNP type molecular markers associated with two sugarcane feedstock quality traits in an experimental breeding population. Molecular markers information was included in the BLASSO statistical model. The correlation coefficients between measured and predicted breeding values were high for both traits in the calibration set. However, the correlation coefficients of the validation set had rather low magnitudes: 0.147 for FIB% and 0.18 for PC%. In addition, the errors of prediction were two-fold higher for FIB% in comparison with PC% (Table 2).

Table 2. Performance of BLASSO in the calibration and validation sets for FIB% and PC%.

	FIB%	PC%
Calibration		
R_c	0.986	0.981
RMSEC	0.267	0.374
Validation		
R_p	0.147	0.180
RMSEP	1.030	0.550

FIB%: fibre content; PC%: apparent sucrose content; BLASSO: Bayesian least absolute shrinkage selection operator; R_p: correlation coefficient of the validation set; RMSEP: root mean squared error of prediction.

The correlation between measured and predicted breeding values found in the calibration and validation set suggest that our model is overfitted, as it adjusted very well to the training set observations but performed poorly in the validation set (Figure 2).

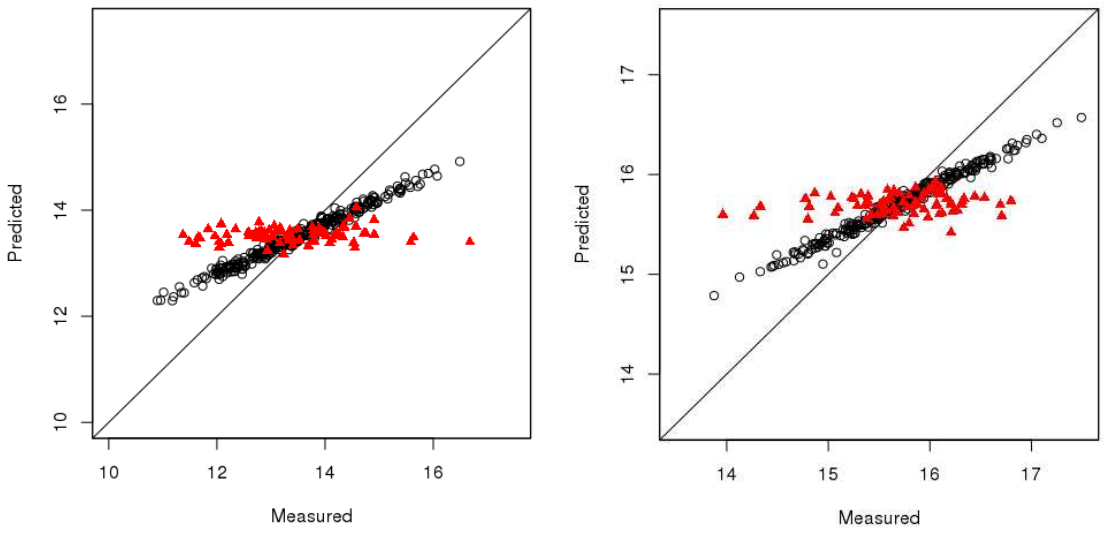


Figure 2. *Left:* Plot of measured versus predicted values from the calibration (empty black circles) and validation set (solid red triangles) for FIB%. *Right:* Plot of measured versus predicted values from the calibration (empty black circles) and validation set (solid red triangles) for PC%.

3.2. Combining genomic and phenomic information

In this section we investigated the incorporation of the NIR spectroscopy data to the SNP array. When we considered the model with the two sets of predictors, the NIR matrix had to be mathematically treated. Thus, we tested several combinations of pretreatments. The performance of BLASSO fitting the model with different combinations of treatment applied in the NIR matrix before the simultaneous analyses is shown in Table 3. The pretreatments that yielded to the best results were Savitzky-Golay smoothing (window: 5; polynomial order: 2), multiplicative scatter correction and mean centre for FIB% and Savitzky-Golay smoothing (window: 5; polynomial order: 2), 1º derivative, multiplicative scatter correction and mean centre for PC% (Table 3). The best models had correlation coefficient values of 0.668 and 0.536, and prediction errors of 0.773 and 0.472 for FIB% and PC%, respectively.

There was a significant increase in the correlation coefficient of prediction values when considering the models developed using the matrix containing only molelular

markers information and the two sets of predictors combined. However, the models for predicting PC% had smaller correlation coefficients than the sole use of NIR spectroscopy data applying either BLASSO or PLS regression. The correlation between measured and predicted breeding values found in the calibration and validation set for FIB% and PC% are shown in Figure 3.

Table 3. Performance of BLASSO models involving SNPs and NIR predictors with different combinations of pretreatments applied in the NIR dataset.

	Pretreatment	R_p	RMSEP
FIB%	Raw	0.661	0.774
	MC	0.614	0.817
	S+MSC+MC	0.668	0.773
	S+D1+MC	0.653	0.790
	S+D1+MSC+MC	0.655	0.783
	S+D2+MC	0.237	1.023
	S+D2+MSC+MC	0.396	0.954
	S+MC	0.596	0.828
PC%	Raw	0.195	0.548
	MC	0.380	0.519
	S+MSC+MC	0.510	0.481
	S+D1+MC	0.522	0.477
	S+D1+MSC+MC	0.536	0.472
	S+D2+MC	0.194	1.025
	S+D2+MSC+MC	0.188	1.025
	S+MC	0.375	0.520

FIB%: fibre content; PC%: apparent sucrose content; MC: mean centre; S: Savitzky-Golay smoothing; MSC: multiplicative scatter correction; D1: 1º derivative; D2: 2º derivative; R_p: correlation coefficient of prediction; RMSEP: root mean squared error of prediction.

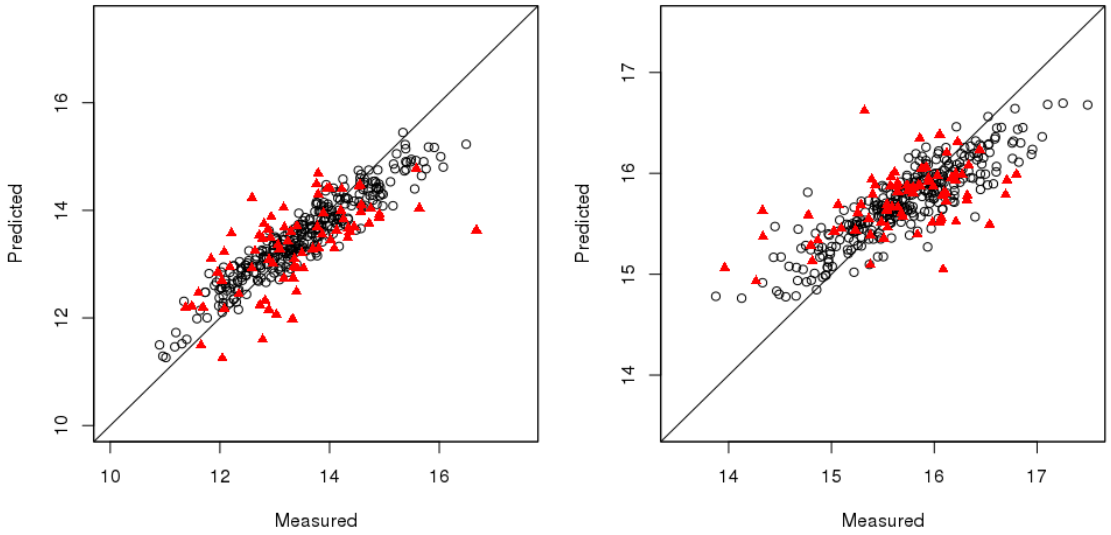


Figure 3. *Left:* Plot of measured versus predicted values from the calibration (black empty circles) and validation (red solid triangles) set for FIB%. *Right:* Plot of measured versus predicted values from the calibration (black empty circles) and validation set (red solid triangles) for PC%.

4. Discussion

The genomic models developed showed poor predictive ability and could only explain a small portion of the genotypic variation of the two traits evaluated. Numerous factors can affect the results of genomic selection. For instance, marker density and the amount of existing linkage disequilibrium (LD; Jannink et al., 2010). De los Campos et al. (2013) affirm that size and the genetic relatedness between training and validation population genotypes is a paramount aspect. The genome structure of the population in which we want to infer breeding values should be similar to the one we used to train our model (Jonas et al., 2013). Moreover, the prediction accuracy of regression methods is affected by the genetic architecture of the trait and different models may yield divergent results (De Los Campos et al., 2013). However, Gouy et al.(2013) applied genomic selection in sugarcane and evaluated different regression models. The authors reported no significant difference between statistical methods for quantitative and qualitative traits.

Zeni Neto et al.(2013) reported that additive and non-additive genetic effects are equally important for the determination of complex traits in sugarcane. Hence, the inclusion of non-additive effects in genomic regression models for sugarcane may improve prediction accuracies. Studies conducted by Viana et al. (2017) and Denis et al.(2013), using simulated data, indicate the increase of efficiency of GS when

accounting for dominance effects. In addition, studies have suggested that the incorporation of pedigree information to genomic models may result in improved accuracies compared to the sole use of molecular markers (Legarra et al., 2009).

The increase of coverage is another highlighted aspect in genomic selection (Sims et al., 2014; Xu et al., 2012). However, Sousa et al. (2019) applied GS in polyploidy hybrids of *Coffea spp.* and suggested that once the optimal number of SNPs is reached, a plateau in terms of the increase in selective accuracy is observed, and then decreases. Yang et al. (2017) reported that for sugarcane and other crops with a complex genome, the quality of sequencing might be more important than a large number of SNPs. Therefore, sequencing depth is paramount to filter low-quality sequence reads (Yang et al., 2017). Seemingly, the greatest bottleneck of the application of genomic selection in sugarcane, can be addressed to the sequencing step. For instance, due to the big and highly complex genome of sugarcane, a polymorphism observed between two individuals may be linked to the absence and presence of the mutation or to different number of copies of each locus.

Cost and efficiency are key drivers to determine the adoption of genomic selection in a breeding program. The main interest in optimizing the selection process is to reduce breeding cycles by performing early selection of elite genotypes. Currently, at the PMGCA, clonal performance is assessed by field-based collection of phenotypic data, which is costly and time-consuming. In addition, each breeding cycle at a sugarcane breeding program takes two years. Clones are evaluated in plant cane, and the selection is made in the first ratoon. At the T1 phase, mass selection is performed. Hence, the selection accuracy is very low, especially regarding low heritability traits (Meuwissen et al., 2013). As a result, at T1, the replacement of these strategies by high-throughput scientifically gathered data is highly desirable, as it could increase selection accuracy. Moreover, genomic selection has the potential to reduce breeding cycle length and increase genetic gains per unit of time, ultimately reducing the time required for cultivar release (Crossa et al., 2017; Heffner et al., 2010).

Considering plant species, studies applying genomic selection have shown promising results (Meuwissen et al., 2013). However, its application for complex traits in crops is still limited. The situation escalates for plants that possess big and complex genomes such as sugarcane.

In this study, we proposed the joint analysis of high-throughput-derived genomic and phenomic information. Our findings are encouraging and suggest that the incorporation of NIR spectra data in and SNP array could increase prediction accuracy of genomic estimated breeding values and thus, save time and reduce costs. However, the NIR models we developed are local, which means that they could only be applied to infer the genetic merit of new unknown sugarcane samples if they have gone through the same preparation protocol. The stage in which we built our model was the T2 phase. Therefore, if we aimed to select sugarcane clones in the T1 phase a novel model should be calibrated for the same conditions and crop developmet stage.

5. Conclusion

In this chapter, we investigated whether the accuracy of genomic selection is improved by combining the NIR spectra matrix and a SNP genotyping matrix into a single regression analysis. The accuracy of genomic selection models was evaluated using the Kennard-Stone algorithm and computing the correlation between the breeding values obtained using phenotypic measurements and breeding values estimated using genomic information. Combining the NIR spectroscopy information to the genomic dataset improved the correlation coefficient estimates for FIB% but not for PC%. The results found in this study suggest that models including NIR spectra-derived data coupled with molecular markers information are able to yield higher predictive ability. Hence, this approach could be used to enhance the efficiency of selection of sugarcane clones by reducing breeding time cycles and thus, increase genetic gains at the PMGCA.

6. References

ALLWRIGHT, M. R.; TAYLOR, G. Molecular Breeding for Improved Second Generation Bioenergy Crops. *Trends in Plant Science*, v. 1, n. January, p. 43–54, 2016.

ARAUS, L J.; E. CAIRNS, J. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, v. 19, n. 1, p. 52–61, 2014.

BARBOSA, M. H. P. *et al.* Genetic improvement of sugar cane for bioenergy : the Brazilian experience in network research with RIDESA. *Crop Breeding and Applied Biotechnology*, v. S2, p. 87–98, 2012.

BERNARDO, R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, v. 48, n. 5, p. 1649–1664, 2008.

BERRY, D. The plant breeding industry after pure line theory: Lessons from the National Institute of Agricultural Botany. *Studies in History and Philosophy of Biological and Biomedical Sciences*, v. 46, n. 1, p. 25–37, 2014.

BOULESTEIX, A. L.; STRIMMER, K. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *oinformaticsBriefings in Bioinformatics*, v. 8, n. 1, p. 32–44, 2007.

BRASILEIRO, B. P. *et al.* Simulated individual best linear unbiased prediction versus mass selection in sugarcane families. *Crop Science*, v. 56, n. 2, p. 570–575, 2016.

BRERETON, R. G. *et al.* Chemometrics in analytical chemistry—part II: modeling, validation, and applications. *Analytical and Bioanalytical Chemistry*, v. 410, n. 26, p. 6691–6704, 2018.

CROSSA, JOSE *et al.* Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. *Genetics*, v. 186, n. 10, p. 713–724, 2010.

CROSSA, JOSÉ *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in plant science*, v. 22, n. 11, p. 961–975, 2017.

DE LOS CAMPOS, G. *et al.* Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, v. 193, n. 2, p. 327–345, 2013.

DENIS, M.; BOUVET, J. M. Efficiency of genomic selection with models including

dominance effect in the context of Eucalyptus breeding. *Tree Genetics and Genomes*, v. 9, n. 1, p. 37–51, 2013.

DESHMUKH, R. *et al.* Integrating omic approaches for abiotic stress tolerance in soybean. *Frontiers in Plant Science*, v. 5, n. June, p. 1–12, 2014.

ENGEL, J. *et al.* Breaking with trends in pre-processing? *Trends in Analytical Chemistry*, v. 50, p. 96–106, 2013.

GALVÃO, R. K. H. *et al.* A method for calibration and validation subset partitioning. *Talanta*, v. 67, n. 4, p. 736–740, 2005.

GARSMEUR, O. *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature Communications*, v. 9, n. 1, p. 1–10, 2018.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, v. 17, n. 6, p. 333–351, 2016.

GOUY, M. *et al.* Experimental assessment of the accuracy of genomic selection in sugarcane. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, v. 126, n. 10, p. 2575–2586, 2013.

HANEGRAAF, M. C.; BIEWINGA, E. E.; VAN DER BIJL, G. Assessing the ecological and economic sustainability of energy crops. *Biomass and Bioenergy*, v. 15, n. 4–5, p. 345–355, 1998.

HEFFNER, E. L. *et al.* Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Science*, v. 50, n. 5, p. 1681–1690, 2010.

HOANG, N. V. *et al.* Potential for Genetic Improvement of Sugarcane as a Source of Biomass for Biofuels. *Frontiers in Bioengineerung and Biotechnology*, v. 3, n. November, p. 1–15, 2015.

HOULE, D.; GOVINDARAJU, D. R.; OMHOLT, S. Phenomics: The next challenge. *Nature Reviews Genetics*, v. 11, n. 12, p. 855–866, 2010.

JAMES, G. *et al. An Introduction to Statistical Learning with Applications in R*. [S.l: s.n.], 2007.

JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: From

theory to practice. *Briefings in Functional Genomics and Proteomics*, v. 9, n. 2, p. 166–177, 2010.

JONAS, E.; DE KONING, D. J. Does genomic selection have a future in plant breeding? *Trends in Biotechnology*, v. 31, n. 9, p. 497–504, 2013.

LANGRIDGE, P.; FLEURY, D. Making the most of "omics" for crop breeding. *Trends in Biotechnology*, v. 29, n. 1, p. 33–40, 2011.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, v. 92, n. 9, p. 4656–4663, 2009.

LI, H.; DURBIN, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, v. 26, n. 5, p. 589–595, 2010.

LOPES, M. L. *et al.* Ethanol production in Brazil: a bridge between science and industry. *Brazilian Journal of Microbiology*, v. 47, p. 64–76, 2016.

MACKAY, I.; OBER, E.; HICKEY, J. GplusE: Beyond genomic selection. *Food and Energy Security*, v. 4, n. 1, p. 25–35, 2015.

MAMMADOV, J. *et al.* SNP markers and their impact on plant breeding. *The Role of Bioinformatics in Agriculture*, v. 2012, p. 387–413, 2014.

MATSUOKA, S.; FERRO, J.; ARRUDA, P. The Brazilian Experience of Sugarcane Ethanol Industry The Brazilian experience of sugarcane ethanol industry. *In Vitro Cellular & Developmental Biology*, v. 45, n. January, p. 372–381, 2009.

MENDU, V. *et al.* Global bioenergy potential from high-lignin agricultural residue. *Proceedings of the National Academy of Sciences*, v. 109, n. 10, p. 4014–4019, 2012.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, v. 157, p. 1819–1829, 2001.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Accelerating Improvement of Livestock with Genomic Selection. *Annual Review ofAnimal Biosciences*, v. 1, p. 221–237, 2013.

MONTES, J. M.; MELCHINGER, A. E.; REIF, J. C. Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science*, v. 12, n. 10, p. 433–436, 2007.

PARK, T.; CASELLA, G. The Bayesian Lasso. *Journal of the American Statistical Association*, v. 103, n. 482, p. 681–686, 2008.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, v. 198, n. 2, p. 483–495, 2014.

PURCELL, D. E. *et al.* Near-infrared spectroscopy for the prediction of disease ratings for fiji leaf gall in sugarcane clones. *Applied Spectroscopy*, v. 63, n. 4, p. 450–457, 2009.

RESENDE, J. F. R. *et al.* Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). *Genetics*, v. 190, n. 4, p. 1503–1510, 2012.

RESENDE, M. D. V. Software Selegen-REML/BLUP: A useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*, v. 16, n. 4, p. 330–339, 2016.

ROGGO, Y.; DUPONCHEL, L.; HUVENNE, J.-P. Quality Evaluation of Sugar Beet (Beta vulgaris) by Near-Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry*, v. 52, n. 5, p. 1055–1061, 2004.

ROQUE, J. V.; DIAS, L. A. S.; TEÓFILO, R. F. Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies. *Jorunal of the Brazilian Chemistry Society*, v. 28, n. 8, p. 1506–1516, 2017.

SANT' ANNA, I. DE C. *et al.* Multigenerational prediction of genetic values using genome-enabled prediction. *PLoS ONE*, v. 14, n. 1, p. 1–14, 2019.

SCHLÖTTERER, C. The evolution of molecular markers. *Nature Reviews Genetics*, v. 5, n. January, p. 63–69, 2004.

SILVEIRA, F. G. DA *et al.* The optimal number of partial least squares components in genomic selection for pork pH. *Ciência Rural*, v. 47, n. 1, p. 1–5, 2016.

SIMS, D. *et al.* Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, v. 15, n. 2, p. 121–132, 2014.

SINGH, A. *et al.* Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*, v. 21, n. 2, p. 110–124, 2016. Disponível em: <http://dx.doi.org/10.1016/j.tplants.2015.10.015>.

SOUSA, T. V. *et al.* Early Selection Enabled by the Implementation of Genomic

Selection in Coffea arabica Breeding. *Frontiers in Plant Science*, v. 9, n. January, p. 1–12, 2019.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. *J Royal Statist Soc B*, v. 58, n. 1, p. 267–288, 1996.

USAI, M. G.; GODDARD, M. E.; HAYES, B. J. LASSO with cross-validation for genomic selection. *Genetics Research*, v. 91, n. 6, p. 427–436, 2009.

VIANA, J. M. S.; PIEPHO, H.-P.; SILVA, F. F. E. Quantitative genetics theory for genomic selection and efficiency of genotypic value prediction in open-pollinated populations. *Scientia Agricola*, v. 74, n. 1, p. 41–50, 2017.

XAVIER, A. *et al.* Walking through the statistical black boxes of plant breeding. *Theoretical and Applied Genetics*, v. 129, n. 10, p. 1933–1949, 2016.

XU, Y. *et al.* Whole-genome strategies for marker-assisted plant breeding. *Molecular Breeding*, v. 29, n. 4, p. 833–854, 2012.

YANG, X. *et al.* Mining sequence variations in representative polyploid sugarcane germplasm accessions. *BMC Genomics*, v. 18, n. 1, p. 1–16, 2017.

ZENI NETO, H. *et al.* Selection of families and parents of sugarcane (Saccharum spp.) through mixed models by joint analysis of two harvests. *Euphytica*, v. 193, n. 3, p. 391–408, 2013.