JAQUELINE GONÇALVES FERNANDES

PREDIÇÃO FENOTÍPICA EM CANA-DE-AÇÚCAR VIA MODELOS MULTIVARIADOS COM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2016

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

T

F363p
2016

Fernandes, Jaqueline Gonçalves, 19-
Predição fenotípica em cana-de-açúcar via modelos multivariados com dados de espectroscopia no infravermelho próximo / Jaqueline Gonçalves Fernandes. - Viçosa, MG, 2016.
xi, 46f. : il. ; 29 cm.

Orientador : Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.43-46.

1. Análise multivariada. 2. Cana-de-açúcar - Fenótipos - Métodos estatísticos. 3. Espectroscopia no infravermelho . I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.535

JAQUELINE GONÇALVES FERNANDES

PREDIÇÃO FENOTÍPICA EM CANA-DE-AÇÚCAR VIA MODELOS MULTIVARIADOS COM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

APROVADA: 24 de fevereiro de 2016.

Reinaldo Francisco Teófilo
(Coorientador)

Bruno Portela Brasileiro

Luiz Alexandre Peternelli
(Orientador)

À Nossa Senhora da Imaculada Conceição

AGRADECIMENTOS

Agradeço a Deus pela vida, pela saúde e pela proteção.

Aos meus pais Osvaldo e Mercês pelo amor incondicional e todo apoio. À minha irmã Janaina pelo incentivo em iniciar o Mestrado e ao meu namorado Edimar pelo carinho e companhia.

Ao professor Luiz Alexandre Peternelli por ser um orientador sempre presente, por todo apoio e motivação.

Ao professor Reinaldo Francisco Teófilo pelos ensinamentos e apoio.

Ao professor Bruno Portela Brasileiro por todo o apoio durante a realização desse trabalho.

Ao professor Márcio Barbosa pela disponibilidade em esclarecer dúvidas.

A todos os professores do departamento de Estatística pelos ensinamentos e pela amizade.

À Carla e a Anita por estarem sempre dispostas a ajudar.

A todos os colegas do departamento de Estatística, aos colegas do departamento de Química e a colega Cristina Dias pela companhia durante os estudos.

À CAPES pela concessão da bolsa de estudo.

À Universidade Federal de Viçosa pela oportunidade de realizar esse trabalho e por todo o aprendizado.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

FERNANDES, Jaqueline Gonçalves. Universidade Federal de Viçosa, fevereiro de 2016. **PREDIÇÃO FENOTÍPICA EM CANA-DE-AÇÚCAR VIA MODELOS MULTIVARIADOS COM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO.** Orientador: Luiz Alexandre Peternelli. Coorientadores: Reinaldo Francisco Teófilo e Ana Carolina Campana Nascimento.

A produção da cana-de-açúcar desempenha papel fundamental na economia do país. Para o desenvolvimento de variedades que atendam as necessidades atuais e também as necessidades futuras é essencial buscar métodos de fenotipagem que proporcionem maior facilidade de utilização, além de rapidez, exatidão e consistência. Visando contribuir para o desenvolvimento de novas estratégias de fenotipagem, esse trabalho teve como objetivo principal construir modelos de predição fenotípica utilizando calibração multivariada. Foram construídos modelos empregando regressão por quadrados mínimos parciais (PLS), regressão por componentes principais (PCR), *Ridge Regression* (RR-BLUP) e Lasso bayesiano (BLASSO) a partir de dados obtidos com espectroscopia na região infravermelho próximo (NIR) em diferentes tipos de amostras de cana-de-açúcar. Esses modelos foram construídos com o objetivo de predizer o teor de fibra (FIB), o teor de sacarose aparente (PC) e o teor de lignina (LIG). O conjunto de calibração foi composto por 166 clones e o de previsão por 20 clones. Os valores de FIB, PC e LIG variaram respectivamente de 8,36% a 22,53%, 1,78% a 16,89% e 13,79% a 21,08%. Os modelos RR-BLUP e BLASSO apresentaram coeficientes de correlação entre 0,70 e 0,91, valores superiores ou iguais aos dos modelos PLS, que por sua vez foram superiores aos dos modelos obtidos por PCR. Para predição de PC e FIB é aconselhável utilizar amostras de colmo devido ao maior poder preditivo além de ser mais viável devido à maior praticidade quando comparado com as amostras de bagaço. Foi possível construir um modelo eficiente para predizer LIG utilizando amostras de bagaço seco. Todos os modelos escolhidos apresentaram bom desempenho para ranquear os melhores clones de acordo com os caracteres em estudo, apresentando medidas elevadas de acurácia, medidas pequenas da taxa de falso positivo e boa precisão.

ABSTRACT

FERNANDES, Jaqueline Gonçalves. Universidade Federal de Viçosa, February, 2016. **PHENOTYPIC PREDICTION IN SUGARCANE USING MULTIVARIATE MODELS WITH DATA FROM NEAR INFRARED SPECTROSCOPY.** . Adviser: Luiz Alexandre Peternelli. Co-advisers: Reinaldo Francisco Teófilo and Ana Carolina Campana Nascimento.

Sugar cane production plays an important role in the economy of a country. In order to develop varieties that meet the current and the future needs, it is essential to find phenotyping methods that are fast, exact, consistent and easy to be used. Aiming at contributing to the development of new phenotyping strategies, the objective of this study was to build models for phenotypic predictions using multivariate calibration. Models using regression by partial least squares (PLS), principal component regression (PCR), Ridge Regression (RR-BLUP) and Bayesian Lasso (BLASSO) were built, using data obtained with Near-Infrared Spectroscopy (NIRS) from different types of sugarcane samples. These models were constructed in order to predict the fiber (FIB), the apparent sucrose (PC) and lignin (LIG) contents. The calibration group was composed of 166 clones, and the prediction group was composed of 20 clones. The amounts of FIB ranged from 8.36% to 22.53%, from 1.78% to 16.89% for PC and from 13.79% to 21.08% for LIG. The RR-BLUP and BLASSO models showed correlation coefficients between 0.70 and 0.91. Those values are higher than or equal to the PLS models, which in turn were higher than those of the models obtained by PCR. For PC and FIB predictions, it is recommended to use cane stalk samples due to its greater predictive power and feasibility when compared to the bagasse samples. It was possible to build an efficient model to predict LIG using dry sugar cane bagasse samples. All selected models showed good performance when ranking the best clones according to the characters studied. Furthermore, the models presented high accuracy measurements, small false positive rate measurements and good accuracy.

1. INTRODUÇÃO

O interesse em buscar fontes de energia renováveis e sustentáveis cresce a cada dia devido, principalmente, à preocupação com a preservação ambiental já que essas novas fontes podem auxiliar na diminuição de CO_2 na atmosfera (KHESHGI, 1999).

Os biocombustíveis têm como objetivo principal substituir os combustíveis fósseis, permitindo assim, diminuir a quantidade de gases que provocam o efeito estufa (SANTOS, 2012). Nesse sentido o etanol da cana-de-açúcar apresenta indicadores ambientais positivos e é uma boa alternativa na substituição dos combustíveis fósseis (CANILHA, 2009).

Sabendo disso, a procura por biocombustíveis no mundo todo tem se tornado cada dia maior e atualmente, o único combustível que pode conseguir atender essa demanda é o etanol proveniente do caldo da cana-de-açúcar, denominado etanol de primeira geração (SANTOS, 2012). Porém, estudos estão sendo realizados para que seja possível aproveitar também os componentes lignocelulósicos da planta e assim poder produzir etanol e outros biocombustíveis a partir do bagaço da cana-de-açúcar, o etanol de segunda geração (ZHENG, 2009).

Portanto, os países produtores de cana-de-açúcar têm procurado cada vez mais investir no melhoramento genético em virtude da sua importância para obter novas variedades como fontes de energia. Além disso, diversos estudos vêm sendo realizados para que o cultivo da cana-de-açúcar também possa expandir para novas fronteiras agrícolas, mesmo em regiões que não possuem tradição no cultivo da cultura (SANTCHURM, 2012; SILVEIRA et al., 2015).

Os métodos usuais utilizados na avaliação fenotípica pelos programas de melhoramento genético da cana-de-açúcar apresentam custo elevado, necessitam de um tempo considerável para execução e por isso, muitas vezes, para algumas características, o processo de fenotipagem torna-se inviável. Dessa forma, é fundamental o desenvolvimento de novas estratégias de fenotipagem, a exemplo do uso da espectroscopia no infravermelho próximo (NIR) associada a métodos estatísticos multivariados na predição de caracteres de interesse para os melhoristas na seleção de novas variedades de cana-de-

açúcar. O NIR apresenta grande possibilidade de aplicação devido a sua facilidade de utilização, rapidez, exatidão e não geração de resíduos (VALDERRAMA et al., 2007; MORGANO et al., 2005). Além de ser um método não destrutivo, não é necessário o uso de reagentes na predição dos caracteres (OZAKI et al., 2007; BLANCO e VILLARROYA, 2002).

Existem diversas aplicações utilizando dados de NIR em diferentes áreas do conhecimento, como na predição do teor de açúcar (NAWI, 2013) e de lignina (ASSIS, 2010) em cana-de-açúcar, na predição do teor de açúcar em batatas (DONG e SUN, 2014), na predição de proteína, matéria seca e fósforo em grãos de milho (GONTIJO NETO, 2009) e na determinação da qualidade da madeira (DE MUÑIZ, 2012). Além disso, o NIR também foi utilizado para determinar o teor de diferentes nutrientes em adubos líquidos e sólidos (YE, 2005), além das análises de matéria orgânica e fração argila em solos e teores foliares de silício e nitrogênio em cana-de-açúcar (SANTOS, 2010).

A modelagem NIR é feita, principalmente, utilizando calibração multivariada que se resume a aplicar regressão por quadrados mínimos parciais (PLS) ou a regressão por componentes principais (PCR) após a realização das transformações e dos pré-processamentos necessários nos dados coletados (FERREIRA, 2015).

Analisando o formato dos dados obtidos através do NIR e comparando com os dados moleculares usados na seleção genômica (*GS - genomic selection*) é possível identificar uma semelhança estrutural, ou seja, as linhas da matriz **X** correspondem aos indivíduos em ambos os casos enquanto que as colunas correspondem aos valores de número de onda *j* para os dados do NIR. Para os dados da seleção genômica cada coluna corresponde ao valor atribuído a cada SNP (do inglês *Single Nucleotide Polimorphism)*. Dessa forma a estrutura dos dados usada na GS parece ser apropriada para uso com dados de NIR e, portanto, nesse estudo também foi utilizado o modelo *Ridge Regression* (RR-BLUP) (MEUWISSEN et al., 2001) e o modelo Lasso bayesiano (BLASSO) (PARK e CASELLA, 2008).

O principal objetivo desse trabalho foi construir modelos de regressão multivariada empregando os métodos (PLS e PCR) e métodos usados na seleção genômica (RR-BLUP e BLASSO). A partir desses modelos foram realizadas predições fenotípicas de caracteres da cana-de-açúcar de interesse

dos programas de melhoramento genético, mas de difícil avaliação em larga escala. Os objetivos específicos são: (1) avaliar diferentes pré-tratamentos nos dados gerados de espectroscopia NIR e (2) avaliar a acurácia de diferentes modelos na predição fenotípica de diferentes caracteres da cana-de-açúcar.

2. REVISÃO BIBLIOGRÁFICA

2.1. Cana-de-açúcar

No passado, a cana-de-açúcar tinha como finalidade apenas a produção do açúcar mascavo. No entanto, com o passar dos anos começaram a surgir diferentes tipos de açúcares e também o etanol que pode ser classificado com biocombustível avançado já que sua produtividade agrícola é alta e existem centenas de variedades de cana-de-açúcar disponíveis para o agricultor (RODRIGUES, 2011).

A produção da cana-de-açúcar desempenha um papel relevante para a economia brasileira já que o país é o maior exportador de açúcar e etanol do mundo. Além disso, a cada dia o país conquista mais o mercado externo com o uso de biocombustíveis como fonte energética (MAPA, 2016).

A estimativa para produção de cana-de-açúcar para a safra 2015/16 cresceu 3,8% em relação à safra anterior, podendo chegar a 658,7 milhões de toneladas (CONAB, 2016). Mesmo com esse aumento na produção a previsão é que a produção de açúcar deve ser menor que na safra passada já que a produção de etanol tende a aumentar. É esperado um crescimento de 7,4% na produção do etanol hidratado que é utilizado nos veículos *flexfuel* (CONAB, 2016).

2.2. Fibra

A fibra é um material complexo, constituído de 32% a 48% de celulose, 19% a 24% de hemicelulose, 23% a 32% de lignina e uma pequena quantidade de cinzas e extrativos (ROCHA et al., 2012; SOUZA et al., 2013).

A fibra é a parte da cana-de-açúcar responsável pela sustentação da planta, além de desempenhar um importante papel na formação dos órgãos de

condução de seiva e estocagem de caldo. A análise da cana-de-açúcar limpa e com os colmos inteiros pelo método da prensa com secagem do bagaço acarreta num teor de fibra variando entre 9% e 15% (FERNANDES, 2000).

Devido ao grande potencial da cana-de-açúcar para a produção de eletricidade e também de diversos produtos de alto valor agregado, a caracterização dos componentes da fibra é indispensável para a obtenção do máximo rendimento e aproveitamento da biomassa.

A Rede Interuniversitária para o Desenvolvimento do Setor Sucro Energético (RIDESA) iniciou um programa de hibridação envolvendo acessos de *Saccharum spontaneum, Saccharum robustum* e cultivares comerciais, visando o desenvolvimento de clones com teores de fibra acima de 17% e que mantenha os atuais 13% de sacarose (SILVEIRA et al., 2015).

2.3. Sacarose

Durante o processo de fotossíntese a cana-de-açúcar produz uma substância de reserva energética, denominada sacarose (MATSUOKA et al., 2010). Por isso, no início dos programas de melhoramento genético de cana-de-açúcar a atenção dos melhoristas recai exclusivamente no desenvolvimento de variedades com elevado teor de sacarose.

O percentual de sacarose presente em toda a cana-de-açúcar é denominado de Pol%cana (PC), isto é, a sacarose aparente presente no caldo e também na biomassa (FERNANDES, 2000).

2.4. Lignina

A lignina é um polímero tridimensional, amorfo, heterogêneo e altamente ramificado. Essa substância é uma das mais abundantes entre as que estão presentes na biosfera (SILVA, 2011). Ela é responsável pela rigidez estrutural das plantas (RODRIGUES, 2011).

Dos componentes da fibra, a lignina está mais associada à produção de energia, por representar um dos maiores estoques de carbono da natureza (DE GORTER, 2010) e devido ao seu maior poder calorifico em relação aos demais componentes da fibra: celulose e hemicelulose (LOUREIRO, et al., 2011).

Entretanto, quando a intenção é a produção de etanol celulósico é preferível utilizar variedades com alto conteúdo de celulose e hemicelulose e baixo teor de lignina (TEW e COBIL, 2008).

2.5. Espectroscopia no Infravermelho Próximo (NIR)

A região do infravermelho próximo (*Near infrared* – NIR) foi descoberta no ano 1800 pelo astrônomo e músico inglês, Frederick William Herschel. No entanto, essa radiação demorou a despertar o interesse dos cientistas. Apenas em 1950, com o avanço da tecnologia e a construção de equipamentos sofisticados é que o infravermelho passou a ser utilizado, principalmente pelos químicos. No Brasil, o uso do NIR foi iniciado no ano de 1991 (PASQUINI, 2003).

O espectro NIR é formado a partir da transformação da energia de radiação em energia mecânica na qual faz com que os átomos de uma determinada molécula se desloquem com uma frequência diferente dependendo da força de ligação que existe entre eles (PASQUINI, 2003). As informações encontradas nesses espectros têm sido muito exploradas para fins analíticos em diferentes áreas do conhecimento.

O instrumento utilizado para medir esses espectros é chamado de espectrofotômetro que pode ser construído com componentes óticos utilizados para instrumentos UV-Visível e com detectores, normalmente, baseados em silício, PbS e materiais fotocondutores InGaAs (PASQUINI, 2003).

A utilização do NIR é muito vantajosa dado que a obtenção dos espectros é rápida, sendo possível analisar as informações em tempo real, além disso, é um método não destrutivo, não invasivo e que possibilita alcançar resultados precisos. A técnica também reduz os custos de análise devido a possibilidade de serem feitas automatizações, proporcionando uma maior produção, além de dispensar o uso de reagente na predição de muitas variáveis (OZAKI et al., 2007; BLANCO e VILLARROYA, 2002).

A análise de dados espectroscópicos, o que inclui o NIR, compreende as seguintes etapas: pré-tratamentos, calibração multivariada e validação. Essas etapas serão discutidas a seguir.

2.6. Pré-tratamentos

Antes de realizar as análises, geralmente são necessários alguns pré-tratamentos para corrigir erros aleatórios e erros sistemáticos (PASQUINI, 2003).

O principal objetivo da aplicação das técnicas de pré-tratamentos é diminuir as variações indesejadas das medidas dos dados que podem influenciar negativamente nas conclusões finais, possibilitando assim a análise do conjunto de espectros obtidos com mais eficiência (FERREIRA, 2015).

Não há nenhuma orientação real para selecionar o melhor pré-tratamento já que a escolha da metodologia dependerá da aplicação, pois nenhum método é único e ideal para todas as situações (FEUDALE, 2002).

Inicialmente é construída uma matriz, denominada matriz **X**, com i linhas e j colunas, com todos os espectros obtidos. Cada linha da matriz **X** corresponde a uma amostra e cada coluna dessa matriz corresponde a uma variável que, neste caso, são os números de onda, dados em cm^{-1}.

Quando o pré-tratamento é realizado nas linhas da matriz **X**, ele é chamado de transformação e no caso do pré-tratamento aplicado nas colunas dessa matriz de dados, ele é chamado de pré-processamento. Normalmente são testados vários métodos para ter certeza de que o pré-tratamento escolhido é o mais adequado (FERREIRA, 2015).

2.6.1. Transformações

Correspondem aos pré-tratamentos feitos nas linhas, isto é, uma ou mais transformações aplicadas nas linhas. Essa seria a primeira etapa na análise de dados NIR. Algumas transformações são apresentadas a seguir.

A Correção Multiplicativa de Sinal (MSC) tem o objetivo de corrigir o efeito da dispersão da luz devido à falta de homogeneidade das amostras ou de corrigir as variações causadas pelas diferenças no percurso óptico das amostras. Como espectro ideal é considerado o espectro médio do conjunto de dados para o qual se deseja realizar a correção da linha base. Utiliza-se uma regressão linear para calcular o coeficiente angular e linear do gráfico entre o espectro ideal e o espectro que vai ser corrigido (FERREIRA, 2015). A

regressão linear é dada por $x_i = a_i + \bar{x}b_i$. Assim, a expressão para a correção é: $x_{i(msc)} = \frac{x_i - a_i}{b_i}$.

A normalização é usada para corrigir a linha de base. Cada uma das variáveis de uma dada amostra i é dividida por um fator de normalização, isto é, pela norma da amostra i representada por $||x_i||$. O resultado é que todas as amostras estarão numa mesma escala. As normas mais utilizadas são: norma superior ou norma infinita $||x_i||_\infty = \max|x_{ij}|$, norma I_1 dada por $||x_i||_1 = \sum_{j=1}^{J}|x_{ij}|$ e norma I_2,ou norma euclidiana, dada por $||x_i||_2 = \sqrt{\sum_{j=1}^{J} x_{ij}^2}$. A partir de uma dessas normas será calculado o valor normalizado: $x_{ij(n)} = \frac{x_{ij}}{||x||_i}$ (FERREIRA, 2015).

O alisamento é utilizado como método de compressão dos dados e para evidenciar picos que estão ocultos devido a presença de ruídos (FERREIRA, 2015). Existem diferentes tipos de alisamento, um dos mais aplicados é o *Savitzky-Golay* que consiste em ajustar um polinômio e posteriormente substituir os pontos originais pelo ponto médio que é ponderado a partir dos coeficientes do polinômio.

A primeira derivada tem o objetivo de corrigir o deslocamento da linha de base e a **segunda derivada** tem o objetivo de corrigir a inclinação da linha de base (FERREIRA, 2015).

A correção da linha de base (baseline), assim como a derivada, é utilizada para corrigir a linha de base através de uma determinada função. Dessa forma é possível eliminar tanto o deslocamento da linha de base quanto a inclinação (FERREIRA, 2015).

2.6.2. Pré-processamentos

Correspondem aos pré-tratamentos feitos nas colunas. Alguns pré-processamentos são apresentados a seguir.

A centragem na média, também conhecida como remoção da média, tem o objetivo de dar maior importância à distância dos pontos em relação ao valor médio, eliminando dos dados o valor da intensidade de cada variável (NAES, et al., 2002). Se x_{ij} representa o *ij*-ésimo elemento da matriz **X**, então a

centragem dos dados na média pode ser realizada utilizando a seguinte formula: $x_{ij(c)} = x_{ij} - \overline{x_j}$. Em alguns casos não foi necessário centrar os dados na média. Para explicar esse fato foi utilizada a análise de componentes principais (MINGOTTI, 2007).

O escalamento pela variância coloca todas as variáveis em uma mesma escala quando tais variáveis apresentam diferentes unidades entre si ou quando a faixa de variação dos dados é grande (TEÓFILO, 2007). Esse pré-processamento pode ser representado por: $x_{ij(ev)} = \frac{x_{ij}}{s_j}$, em que $s_j = \sqrt{\frac{\sum_{i=1}^{I}(x_{ij}-\bar{x})}{I-1}}$.

O autoescalamento é a realização simultânea dos pré-processamentos centrar na média e escalar pela variância. Sua formula é bastante simples: $x_{ij(a)} = \frac{(x_{ij}-\bar{x}_j)}{s_j}$ (FERREIRA, 2015).

2.7. Calibração Multivariada

Após serem feitos todos os pré-tratamentos necessários, a nova matriz de dados nos permite gerar resultados que são do interesse do pesquisador, isto é, que permite fazer previsões. Esse procedimento recebe o nome de calibração e é dividido em duas etapas: a construção de modelos e a validação desses modelos (FERREIRA, 2015).

A forma mais simples e fácil de construir um modelo de calibração é utilizando o método univariado. Entretanto, esse método é muito restrito já que considera que a resposta é influenciada apenas por uma única variável de interesse enquanto que o método multivariado permite considerar diversas variáveis (FERREIRA, 2015). Além disso, é possível citar outras inúmeras vantagens ao utilizar a calibração multivariada como, por exemplo, a redução no ruído, a possibilidade de construir modelos mesmo com a presença de interferentes após incluí-los no modelo e também controlar amostras anômalas (BRO, 2013).

Para a construção dos modelos de calibração multivariada normalmente são utilizados o método de regressão por componentes principais (PCR) ou o método de regressão por quadrados mínimos parciais (PLS). O primeiro

apresenta algumas desvantagens sobre o segundo já que é matematicamente mais complicado e considera apenas a variância das variáveis independentes (matriz **X**), omitindo as informações presentes nas variáveis dependentes (vetor **y**).

Os modelos de regressão são construídos basicamente resolvendo o seguinte sistema **y=Xb**, em que **b** o vetor de regressão que precisa ser estimado.

Ao construir um modelo utilizando PCR é realizada a decomposição da matriz **X** pelo algoritmo da decomposição dos valores singulares (SVD), o mesmo usado na análise de componentes principais. Dessa forma, a matriz $\mathbf{X}_{(I,J)}$ poderá ser escrita como: $\mathbf{X=USV}^T$, em que $\mathbf{U}_{(I,I)}$ é ortogonal e contêm os vetores singulares do lado esquerdo, $\mathbf{S}_{(I,J)}$ é uma matriz retangular diagonal, na qual estão os valores singulares apresentados em ordem decrescente e relacionados à porcentagem de variância explicada dos dados e, finalmente, $\mathbf{V}_{(J,J)}$ é ortogonal e contém os vetores singulares do lado direito. Em seguida, é realizada a truncagem, ou seja, é feita a seleção e escolha do número ótimo de componentes principais (*H*), alterando as dimensões das matrizes apresentadas anteriormente para $\mathbf{U}_{(I,H)}$, $\mathbf{S}_{(H,H)}$ e $\mathbf{V}_{(J,H)}$. Assim, é possível determinar o vetor de regressão $\mathbf{b}=(\mathbf{V}_{(J,H)}(\mathbf{S}_{(H,H)})^{-1})\mathbf{U}_{(H,I)}{}^T)\mathbf{y}$ após inverter a matriz **X**. Este procedimento é executado usando a pseudo-inversa de Moore-Penronse: (FERREIRA, 2015).

O método PLS é muito usado na modelagem com dados de NIR e seu principal objetivo é diminuir as distâncias das medidas originais. O PLS é construído em uma única etapa, isto é, as informações da matriz **X** e as informações de interesse **y** são consideradas simultaneamente durante a decomposição (FERREIRA, 2015).

A construção do modelo PLS é feita, inicialmente, realizando a compressão de dados, isto é, realizando a escolha das variáveis latentes (H) que contêm informações apresentadas em ordem decrescente de variância (BARLOW et al, 2005). A escolha dessas componentes é feita de forma a maximizar a relação entre **X** e **y**. Em seguida, é aplicado o algoritmo bidiagonal proposto por Manne (MANNE, 2001) que consiste na decomposição da matriz **X**, em outras matrizes, $\mathbf{X=RBA}^T$. E finalmente é calculado o vetor de regressão após realizar a pseudo-inversa $\mathbf{b}=(\mathbf{A}_{(J,H)}(\mathbf{B}_{(H,H)})^{-1}\mathbf{R}_{(H,I)}{}^T)\mathbf{y}$.

Durante a construção do modelo é feito a validação cruzada para determinar o número de variáveis latentes no conjunto de calibração. Essa validação pode ser feita retirando apenas uma amostra (*leave one out*) ou um grupo de amostras do conjunto total de dados e, então, o modelo é construído para que os valores sejam estimados. O processo é repetido até que todos os valores sejam previstos (VALDERRAMA et al., 2007; MORGANO et al., 2008).

Para avaliar os modelos são usados alguns valores como o coeficiente de correlação, o coeficiente de determinação (R^2), a raiz do erro quadrático médio de validação cruzada (*RMSECV*) e a raiz do erro quadrático médio de previsão (*RMSEP*). O modelo pode ser considerado adequado quando os valores de *RMSECV* e de *RMSEP* forem pequenos (TEÓFILO, 2013), isto é, menores que o valor mínimo da variável dependente.

O coeficiente de determinação múltipla (R^2) é calculado a partir da fórmula apresentada abaixo, onde $\bar{y}$ é o valor médio medido e $\bar{\hat{y}}$ é o valor médio predito.

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

Além de ser usado para avaliar o desempenho de regressões, o *RMSE* também é utilizado para escolher o número ótimo de componentes principais para os modelos PCR e PLS (HARALD e TORMOD, 1989). A equação a seguir mostra como é calculado o RMSE, onde y_i é o valor real da amostra i, $\hat{y}_i$é o valor estimado para a amostra *i*, e n é o número total de amostras.

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

2.8. Métodos e modelos estatísticos para predição genômica

2.8.1. *Ridge Regression* (RR-BLUP)

A *Ridge Regression* tem como objetivo diminuir a variância das estimativas dos parâmetros a partir de alterações feitas nos coeficientes, nos quais são aproximados para o valor zero de forma que seus efeitos não sejam

anulados (WHITTAKER et al, 2000). As estimativas dos coeficientes são obtidas por meio da equação:

$$\hat{\beta}^{ridge} = argmin_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \tilde{\lambda}\sum_{j=1}^{p}\beta_j^2\right\}$$

onde, y_i é o valor fenotípico do indivíduo i, x_i é a i-ésima linha da matriz de variáveis independentes, β é o vetor dos coeficientes de regressão correspondentes, $\sum_{j=1}^{p}|\beta_j|$ é a penalização, $\tilde{\lambda}$ é o parâmetro de ajuste ou regularização e β_jé a estimativa do j-ésimo número de onda. O parâmetro $\tilde{\lambda}$ controla o balanço entre o ajuste do modelo, medido pela soma de quadrado de resíduos $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ e a complexidade do modelo, medida pela soma de quadrados dos efeitos das variáveis independentes $\sum_{j=1}^{p}\beta_j^2$ (HASTIE et al, 2009; PÉREZ e DE LOS CAMPOS, 2014). Quanto maior o valor de $\tilde{\lambda}$ maior será o encolhimento dos coeficientes (RESENDE et al, 2011).

O método RR-BLUP apresenta a vantagem de prevenir problemas de multicolinearidade. No entanto, é um método complexo já que mantém todas as covariáveis (RESENDE et al, 2011).

.

2.8.2. Lasso bayesiano (BLASSO)

O método bayesiano BLASSO assume uma distribuição a priori dupla exponencial para os efeitos de cada variável e faz com que alguns coeficientes muito pequenos sejam anulados além de selecionar variáveis (RESENDE et al, 2011). As estimativas dos coeficientes são obtidas a partir da seguinte fórmula:

$$\hat{\beta} = min\left\{\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_L\sum_{j=1}^{p}|\beta_j|\right\}$$

onde, y_i é o valor fenotípico do indivíduo i, x_i é a i-ésima linha da matriz de variáveis independentes, β é o vetor dos coeficientes de regressão correspondentes, $\lambda_L\sum_{j=1}^{p}|\beta_j|$ é o parâmetro de regularização que mantém as covariáveis mais significativas e remove as demais, e β_jé a estimativa do efeito da j-ésima variável independente (RESENDE et al, 2011).

Para cada variável é estimada uma variância e assim é possível obter um maior encolhimento entre os coeficientes que estão em torno do valor zero e um menor encolhimento dos coeficientes que apresentam valores mais distantes de zero (RESENDE et al, 2011).

3. MATERIAL E MÉTODOS

3.1. Material Vegetal

Após a seleção clonal realizada nas melhores famílias avaliadas na primeira fase de teste (T1) do Programa de Melhoramento Genético da Cana-de-açúcar pertencente à Universidade Federal de Viçosa (PMGCA/UFV), 160 clones foram conduzidos para a segunda fase de teste (T2).

Os clones selecionados são contrastantes em relação aos caracteres teor de fibra (FIB) e teor de sacarose aparente (PC). Esses clones são oriundos de cruzamentos envolvendo genitores com alto teor de sacarose e genitores com elevado teor de fibra, visando à obtenção de cultivares de cana energia com maior teor de fibra e com os mesmos teores de sacarose das cultivares atuais.

O T2 foi instalado em julho de 2014 no delineamento em blocos aumentados (DBA) com duas cultivares como testemunhas (RB867515 e C90-176) e com 18 repetições (blocos). Cada uma das 196 parcelas foi constituída por dois sulcos de 5 m de comprimento, espaçados em 1,4 m. A instalação do experimento foi realizada no Centro de Pesquisa e Melhoramento Genético da Cana-de-Açúcar (CECA) pertencente à Universidade Federal de Viçosa, no município de Oratórios, MG, com latitude 20º25’ S; longitude 42º48’W; altitude 494m; solo LVE.

3.2. Avaliação Fenotípica

Os caracteres teor de fibra (FIB) e teor de sacarose aparente (PC) foram obtidos por meio da análise tecnológica realizada em amostras de 500 g oriundas da moagem de 10 canas por parcela aos 10 meses (maio de 2015)

após o plantio, no início da safra e aos 13 meses após o plantio (agosto de 2015), meio de safra (FERNANDES, 2000).

Em 120 das 196 parcelas, foi realizada a fenotipagem para o teor de lignina (LIG) da fibra via empresa celignis (www.celignis.com) que obteve os fenótipos fazendo uso de modelos de calibração previamente construídos naquela empresa.

Para a obtenção do teor de lignina (LIG) foram usadas amostras do bagaço seco, moído e peneirado oriundo do processo de análise tecnológica descrita acima, tratamento necessário para as análises dos componentes presentes na fibra.

3.3. Obtenção dos espectros

Aproximadamente 200 g do bagaço com caldo, obtidos a partir da moagem de 10 colmos de cada uma das 196 parcelas, foram congelados a -20ºC visando evitar o processo de fermentação e deterioração até a leitura das amostras NIR.

As leituras da transflectância via NIR foram realizadas após 30 dias de armazenamento em um espectrômetro com transformada de Fourier (FT) Agilent 660. De cada amostra descongelada foram usados aproximadamente 3 g para três leituras, a cada leitura de uma mesma amostra o recipiente contendo o bagaço era movimentado para que uma nova leitura fosse realizada em outro ponto da amostra (Figura 1). Em três dias todas as 588 leituras foram realizadas.

As amostras de colmo também foram usadas nas leituras NIR obtidas através do mesmo espectrômetro com transformada de Fourier (FT) Agilent 660. Nesse caso, três entrenós do terço médio de três colmos (nove entrenós) por parcela foram congelados a -20ºC visando evitar a deterioração das amostras. Após 30 dias iniciou-se o descongelamento dos entrenós das 196 amostras de colmo. Cada entrenó foi cortado longitudinalmente e usado para uma única leitura no NIR (Figura 2). Foram necessários sete dias para realizar as 1764 leituras.

Figura 1 - Preparo das amostras contendo bagaço com caldo (a) e realização das leituras via *Near Infrared* (NIR) com o espectrômetro (b).

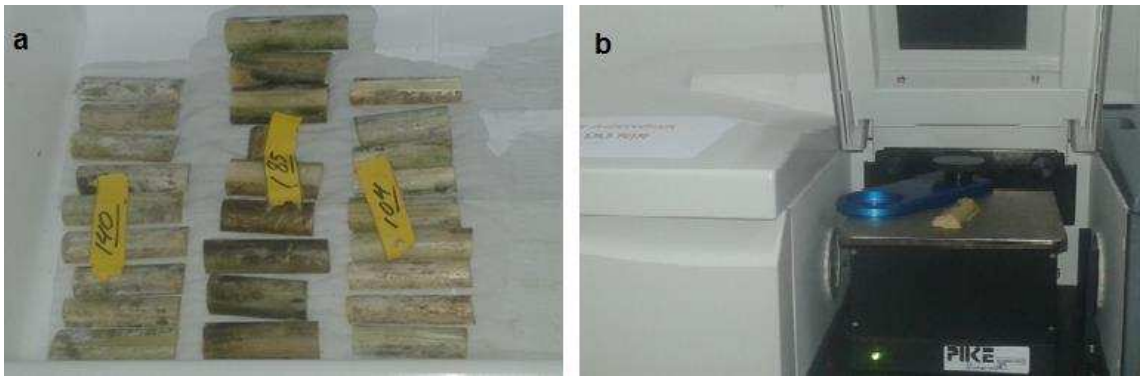

Figura 2 - Preparo das amostras, contendo 9 entrenós de 3 colmos de cada parcela (a) e realização das leituras via Near Infrared (NIR) com o espectrômetro (b).

Para a modelagem do caracter lignina (LIG), aproximadamente, 2 g do bagaço seco e moído, obtido a partir da moagem de 10 colmos de cada uma das 196 parcelas, no estágio de cana planta, foram usadas nas leituras via infravermelho próximo (NIR). De cada amostra foram realizadas três leituras, a cada leitura de uma amostra o recipiente contendo o bagaço era movimentado para que a nova leitura fosse realizada em outro ponto da amostra.

3.4. Análise Estatística

As análises foram realizados por meio do pacote de algoritimo PLS-Toolbox 4.0, executado no *software* Matlab versão 6.0 (*The Mathworks, Natick,* USA).

Os dados analisados são compostos por uma matriz **X**, em que cada linha é um conjunto de informações sobre cada um dos indivíduos e as colunas são as variáveis em estudo que, neste caso, são os números de onda do NIR.

Inicialmente, cada conjunto de dados foi dividido em dois conjuntos utilizando o algoritmo de Kennard e Stone*:* um conjunto de calibração e um conjunto de validação externa. Esse método permite selecionar amostras de forma uniforme (KENNARD e STONE, 1969).

Em seguida foi realizada a validação cruzada *leave-one-out* (VALDERRAMA et al., 2007; MORGANO et al., 2008) nos dados de calibração para escolher a melhor quantidade de variáveis latentes para a construção dos diferentes modelos, tanto pelo método PLS quanto pelo método PCR, nos quais foram testados os pré-tratamentos 1ª e 2ª derivadas, MSC e centrar na média (FERREIRA, 2015). Além disso, foram construídos gráficos a partir da análise de componentes principais (MINGOTT, 2007) para melhor entender o comportamento dos dados.

Foi avaliada ainda a existência de amostras anômalas nos conjuntos de dados através do gráfico de Leverage versus Resíduos estudentizados (MARTENS e NAES, 1989). A Leverage é uma medida que indica o quanto uma determinada amostra influencia no modelo construído. É possível aplicar o teste t supondo que os resíduos seguem uma distribuição normal e se o número de amostras é grande os resíduos estão dentro do intervalo de mais ou menos 2,575 ao nível de 99% de probabilidade (BARTHUS, 2005). Assim, usa-se um intervalo igual a mais ou menos três como ponto crítico para a identificação das amostras anômalas. Os Resíduos estudentizados tem distribuição t-*Student* com $(n-p-1)$ graus de liberdade e são obtidos da regressão original com todas as observações. Além disso, os resíduos estudentizados podem ser usados para verificar se um determinado ponto é aberrante quando comparado com o quantil $t_{\left(1-\frac{\alpha}{2},\ n-p-1\right)}$ correspondente (CORDEIRO e LIMA NETO, 2006).

Finalmente, foi realizada a construção dos modelos escolhidos nos conjuntos de dados de validação externa e então foi possível calcular o RMSEP.

Considerando a semelhança com a estrutura de dados usados na seleção genômica, em que ao invés de número de onda teríamos marcadores moleculares, serão usados, na modelagem, os modelos denominados RR-

BLUP (MEUWISSEN et al., 2001) e o método bayesiano BLASSO (PARK e CASELLA, 2008).

3.5. Avaliação e comparações técnicas

Com a intenção de avaliar a classificação dos clones a partir dos melhores modelos escolhidos nas etapas anteriores foram construídas matrizes de confusão (Tabela 1) para cada uma das variáveis, teor de fibra, teor de sacarose aparente e lignina mostrando as medidas de acurácia, taxa de falso positivo e precisão.

A medida de acurácia se refere à soma do número de clones selecionados e não selecionados, tanto pela classificação dos valores reais quanto pela classificação dos valores medidos, dividido pelo total de clones. A taxa de falso positivo se refere ao número de clones selecionados erroneamente, a partir dos valores preditos, dividido pelo total de selecionados a partir dos valores reais. E precisão é o número de clones selecionados, simultaneamente, pelos valores reais e pelos valores preditos após dividir pelo total de clones selecionados a partir dos valores reais.

A matriz de confusão mostra o número de clones selecionados ou não selecionados a partir das respostas reais versus respostas preditas. As respostas reais são os valores de cada uma das variáveis que inicialmente foram separadas para validação externa e as respostas preditas são os valores estimados para essas respostas a partir do modelo escolhido. Um esquema de como é apresentada a matriz de confusão é exemplificado na Tabela 1.

Tabela 1 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais e a classificação obtida a partir dos valores preditos a partir dos modelos.

Classificação	NS (Preditos)	S (Preditos)
NS (Real)	VN	FP
S (Real)	FN	VP
Acurácia	(VN+VP) / n	
Taxa de falso positivo	FP / (FN + VP)	
Precisão	VP / (FN + VP)	

* S= Clones selecionadas, NS= Clones não selecionadas, VP= Verdadeiro Positivo, FN= Falso Negativo, FP= Falso Positivo, VN= Verdadeiro Negativo, n=número total de clones.

O percentual de clones não selecionados tanto pelos valores medidos quanto pelos valores preditos é representado por VN. FN corresponde ao percentual de clones selecionados a partir dos valores medidos que não foram selecionadas entre os valores preditos. FP indica o percentual de clones que não seriam selecionadas com base nos valores reais, mas que foram selecionadas pelos valores preditos. Finalmente, VN corresponde ao percentual de clones não selecionados nos dois casos.

A matriz de confusão nessa dissertação foi construída para o conjunto de teste supondo a seleção de 50% e 25% do total de clones com as maiores médias para o teor de fibra (FIB), teor de sacarose aparente (PC) e teor de lignina (LIG). Além disso, foi construída uma matriz de confusão para (FIB) supondo a seleção materiais com FIB maior que 17% (SILVEIRA et al., 2015).

4. RESULTADOS E DISCUSSÃO

4.1. Análises espectroscópicas

Os espectros NIR obtidos a partir das amostras de bagaço úmido coletado no início da safra, bagaço úmido coletado no meio da safra, bagaço seco e de colmo estão respetivamente apresentados nas Figuras 3, 4, 5 e 6. Cada conjunto de dados foi separado pelo algoritmo KENNARD e STONE (KENNARD e STONE, 1969) em dois conjuntos: o primeiro se refere aos dados das variáveis independentes utilizadas para a construção dos modelos e o

segundo se refere aos dados das variáveis independentes utilizados para a validação externa do modelo escolhido.

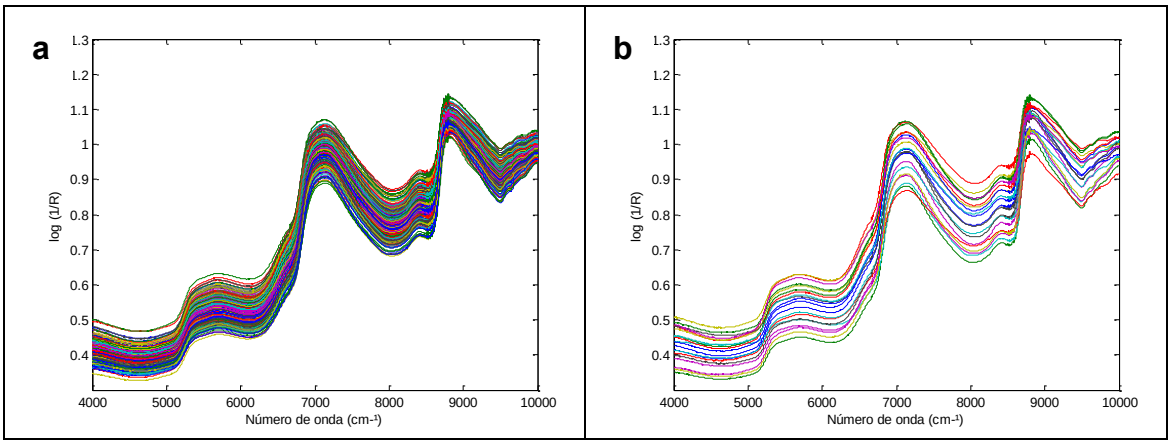

Figura 3 - Espectros NIR obtidos a partir do bagaço úmido congelado coletado no início da safra para os dados de calibração (a) e para os dados de validação (b).

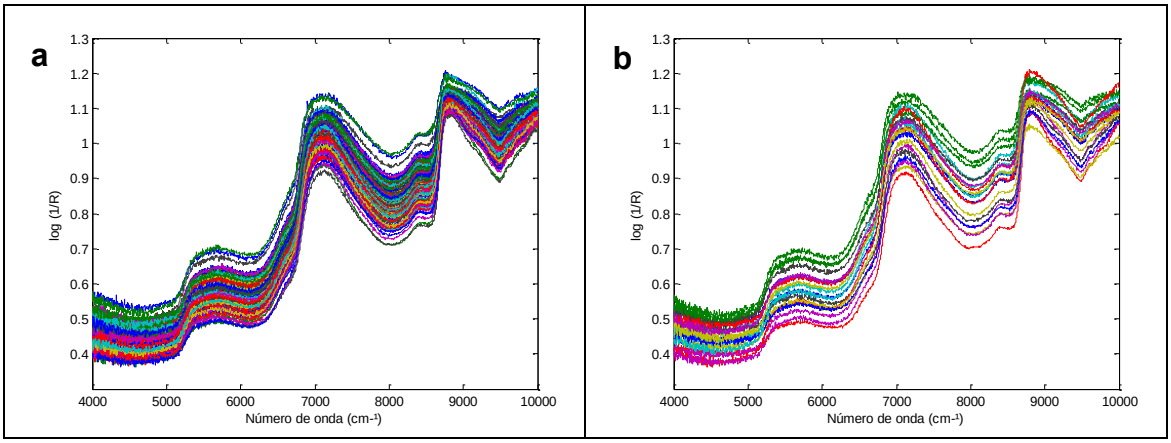

Figura 4 - Espectros NIR obtidos a partir do bagaço úmido congelado coletado no meio da safra para os dados de calibração (a) e para os dados de validação (b).

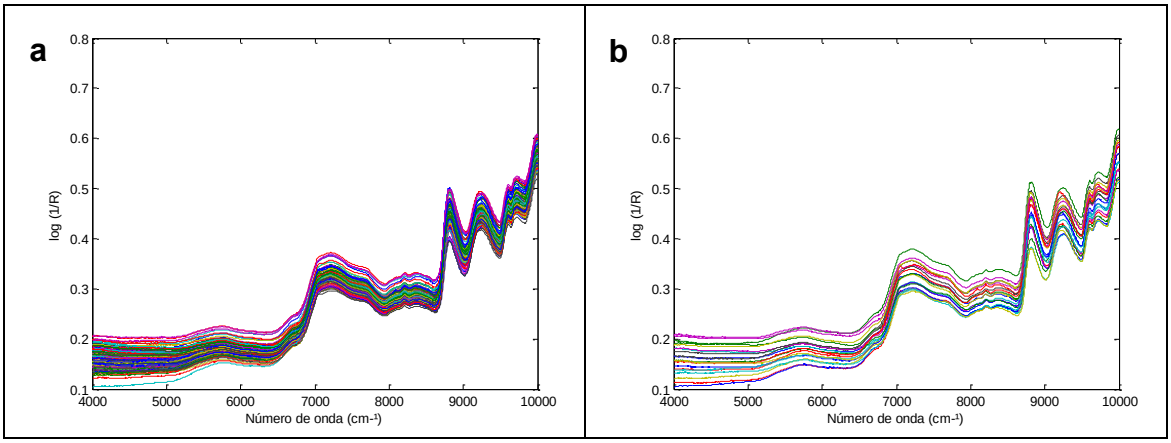

Figura 5 - Espectros NIR obtidos a partir do bagaço seco coletado no meio da safra para os dados de calibração (a) e para os dados de validação (b).

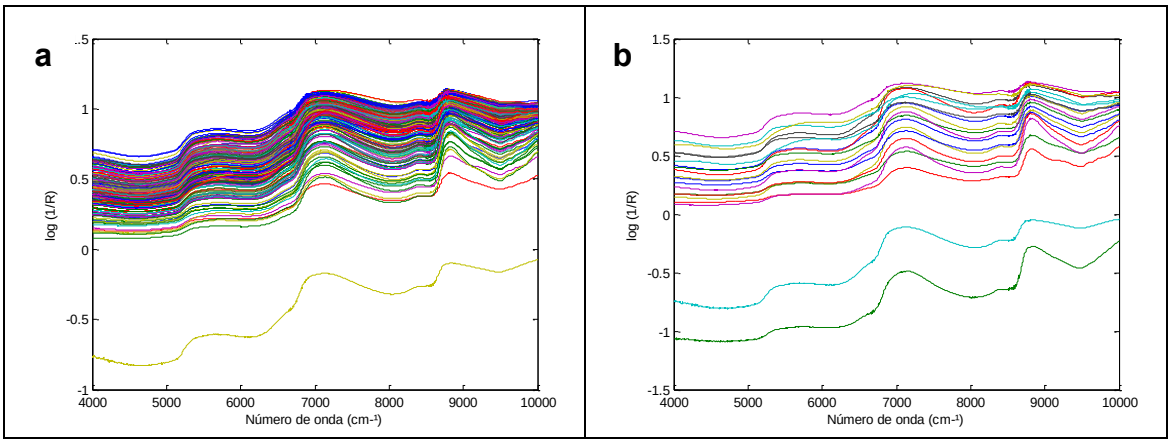

Figura 6 - Espectros NIR obtidos a partir do colmo coletado no meio da safra para os dados de calibração (a) e para os dados de validação (b).

4.2. Teor de fibra (FIB)

O teor de fibra variou entre 8,38% a 19,51% no início da safra. No meio da safra esse teor aumentou, variando entre 9,58% e 22,53% conforme apresentado na Tabela 2 de análises descritivas. A Figura 7 mostra gráficos de boxplot estratificados pelos conjuntos de dados de bagaço congelado coletados no início e no meio da safra e também pelos dados de colmo coletados no meio da safra.

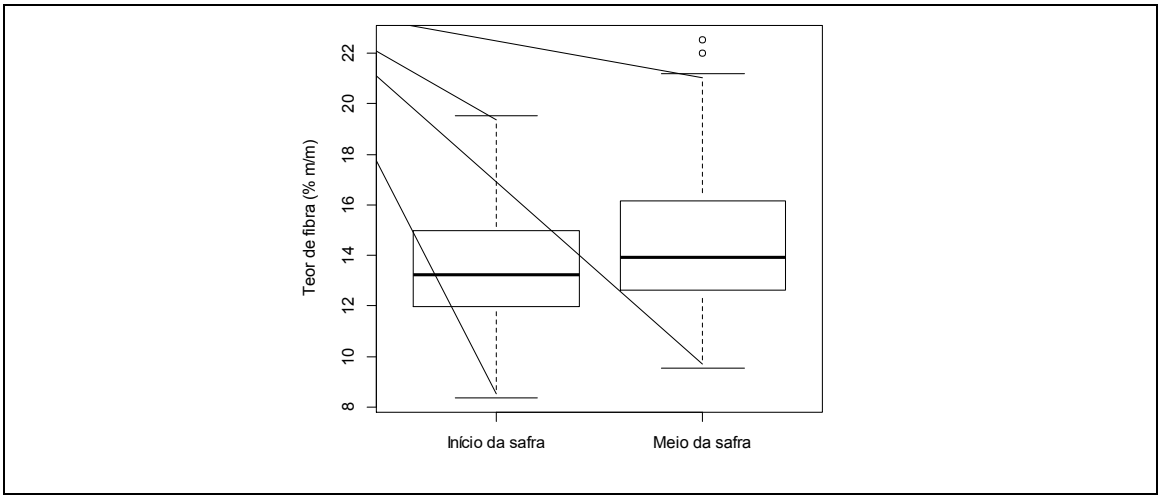

Figura 7 - Boxplots com os valores do teor de fibra (FIB) obtidos a partir da análise tecnológica realizadas em amostras de bagaço congelado coletado no início da safra e em amostras de bagaço congelado coletado no meio da safra.

Tabela 2 - Análise descritiva dos valores do teor de fibra obtidos a partir da análise tecnológica realizadas em amostras de bagaço no início e em amostras de bagaço no meio da safra.

	Início da safra	Meio da safra
Média	13,48	14,44
Mediana	13,23	13,93
Mínimo	8,36	9,56
Máximo	19,52	22,53
Desvio Padrão	2,06	2,42

Com o interesse de predizer esses percentuais de fibra a partir do bagaço e também do colmo da cana-de-açúcar foram escolhidas 20 amostras de cada um dos bancos de dados, a partir do algoritmo KENNARD e STONE (KENNARD e STONE, 1969) que corresponde a aproximadamente 10% do conjunto de dados para realizar a validação externa, isto é, para fazer a previsão. As demais amostras foram usadas para o treinamento dos modelos com diferentes tratamentos como mostrado na Tabela 3.

A escolha do número de variáveis latentes (VL) para a construção dos modelos foi feita a partir do gráfico da raiz do erro quadrático médio de validação cruzada (*RMSECV*) versus VL apresentado na Figura 8.

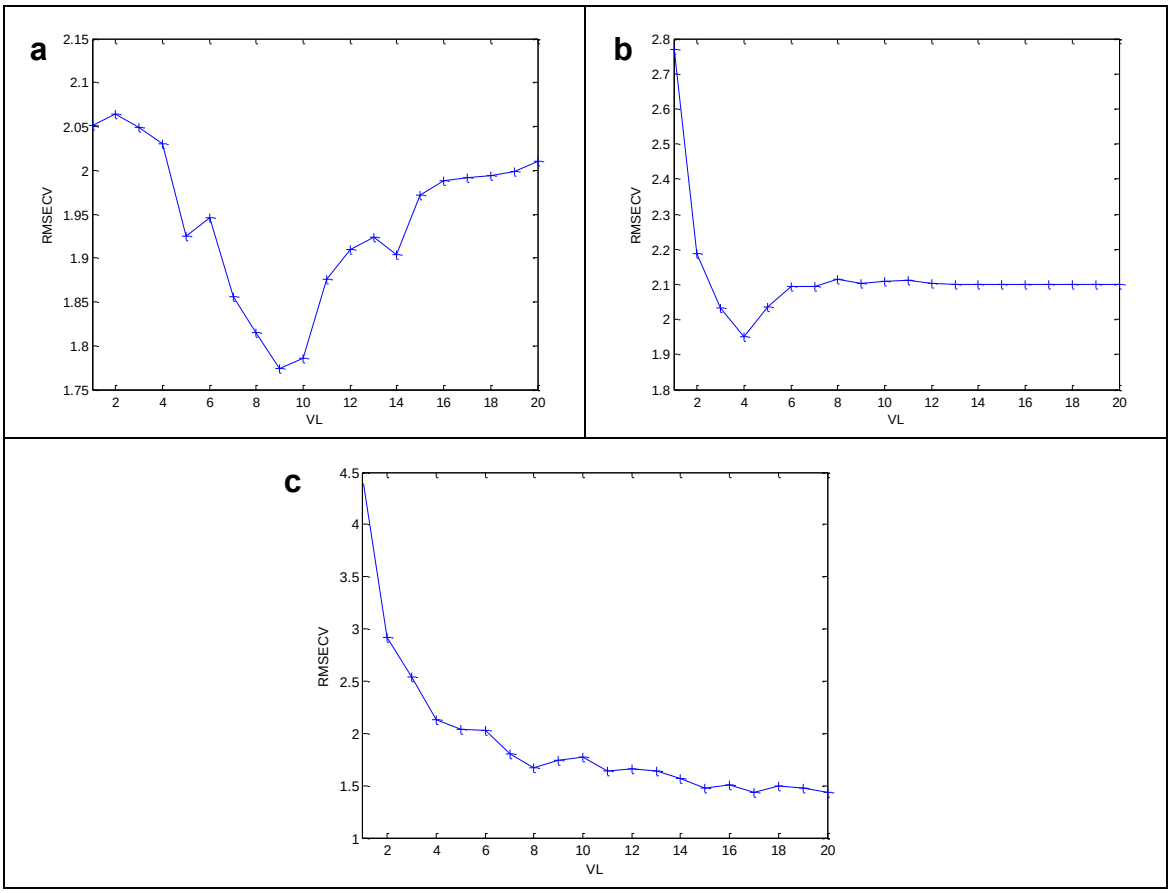

Figura 8 - Número de variáveis latentes (VL) versus a raiz do erro quadrático médio de validação cruzada *(RMSECV)* em dados de bagaço coletados no início da safra após centrar na média (a), em dados de bagaço coletados no meio da safra sem nenhum pré-tratamento (b) e em dados de colmo coletados no meio da safra sem nenhum pré-tratamento (c).

Para cada conjunto de dados foi ajustado um determinado modelo que apresentava o maior valor do coeficiente de determinação e então foi escolhido o menor número de VL correspondente ao menor valor de *RMSECV*.

A Figura 8 mostra que o menor valor de *RMSECV* ocorre quando o número de variáveis latentes foi igual a nove para os dados de bagaço coletados no início da safra e igual a quatro para os dados de bagaço coletados no meio da safra. Para os dados de colmo optou-se por escolher o número de VL igual a oito visto que o valor de *RMSECV* é um dos menores entre os apresentados.

Para escolher o melhor modelo para predição do teor de fibra da cana-de-açúcar a partir de espectros NIR obtidos tanto do bagaço quanto do colmo foram construídos modelos PLS e PCR testando diferentes tratamentos como 1ª e 2ª derivadas, MSC e centrar na média. Os resultados estão apresentados nas Tabelas 3 e 4.

Tabela 3 - Valores de *RMSECV* e de *R²* para diferentes tratamentos em dados de bagaço úmido coletado no início da safra, bagaço úmido coletado no meio da safra e colmo da cana-de-açúcar para predição do teor de fibra utilizando o modelo PLS.

	bagaço do início da safra			bagaço do meio safra			colmo do meio da safra		
Pré-tratamentos	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**
Nenhum	**9**	**1,77**	**0,29**	**4**	**1,95**	**0,34**	**8**	**1,67**	**0,47**
Centrar na Média (CM)	**9**	**1,77**	**0,29**	4	1,97	0,33	8	2,15	0,28
1ª Derivada + CM	9	2,22	0,06	4	2,28	0,21	8	1,78	0,43
2ª Derivada + CM	9	2,42	0,00	4	2,63	0,03	8	1,89	0,34
MSC + CM	9	1,85	0,26	4	2,13	0,27	8	2,11	0,27
MSC + 1ª Derivada + CM	9	2,22	0,06	4	2,27	0,21	8	2,06	0,34
MSC + 2ª Derivada + CM	9	2,41	0,00	4	2,62	0,03	8	1,88	0,35

Tabela 4 - Valores de *RMSECV* e de *R²* para diferentes tratamentos em dados de bagaço úmido coletado no início da safra, bagaço úmido coletado no meio da safra e colmo da cana-de-açúcar para predição do teor de fibra utilizando o modelo PCR.

	bagaço do início da safra			bagaço do meio safra			colmo do meio da safra		
Pré-tratamentos	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**
Nenhum	9	1,98	0,10	4	1,99	0,31	8	2,15	0,21
Centrar na Média (CM)	9	1,97	0,12	4	1,98	0,31	8	2,86	0,07
1ª Derivada + CM	9	1,85	0,21	4	2,01	0,29	8	1,79	0,37
2ª Derivada + CM	9	2,08	0,02	4	2,44	0,23	8	1,71	0,41
MSC + CM	**9**	**1,99**	**0,10**	**4**	**1,94**	**0,34**	**8**	**1,70**	**0,42**
MSC + 1ª Derivada + CM	9	1,85	0,21	4	1,98	0,31	8	1,77	0,38
MSC + 2ª Derivada + CM	9	2,08	0,02	4	2,44	0,22	8	1,74	0,39

A Tabela 3 indica que os dados coletados já podem estar próximos da média visto que os melhores modelos foram construídos a partir dos dados sem aplicar o pré-processamento centrar na média.

As Figuras 9, 10 e 11 mostram os gráficos construídos a partir da análise de componentes principais (PCA). Percebe-se que soma das variâncias do componente principal (PC1) e do componente principal (PC2) não varia muito entre o primeiro gráfico com os dados centrados na média e o segundo gráfico com os dados sem nenhum pré-tratamento. Dessa forma, é possível explicar a escolha dos melhores modelos sem nenhum tipo de pré-tratamento.

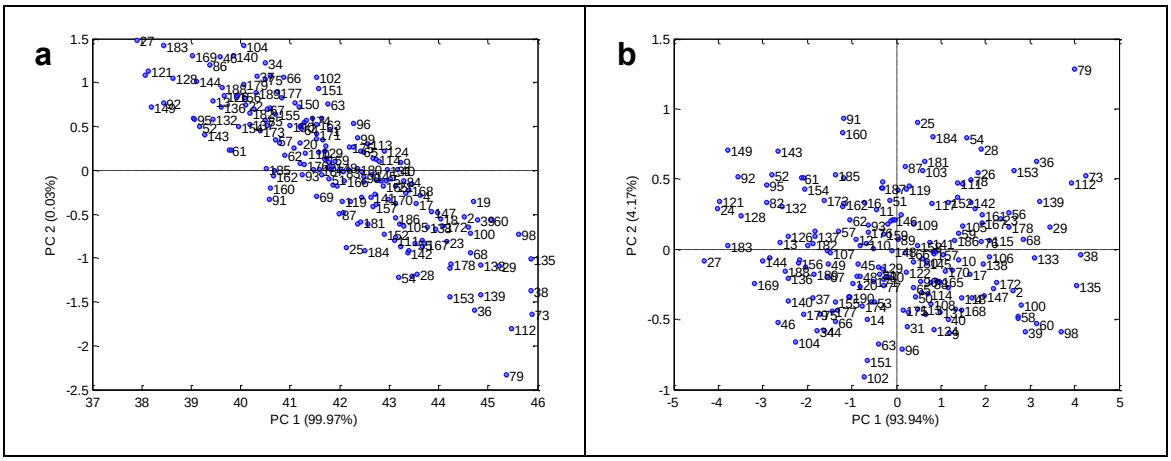

Figura 9 - Gráficos de dispersão a partir dos dois primeiros componentes principais (PC1 e PC2) para os dados de bagaço coletados no início da safra sem nenhum pré-tratamento (a) e para os dados de bagaço coletados no início da safra após serem centrados na média (b).

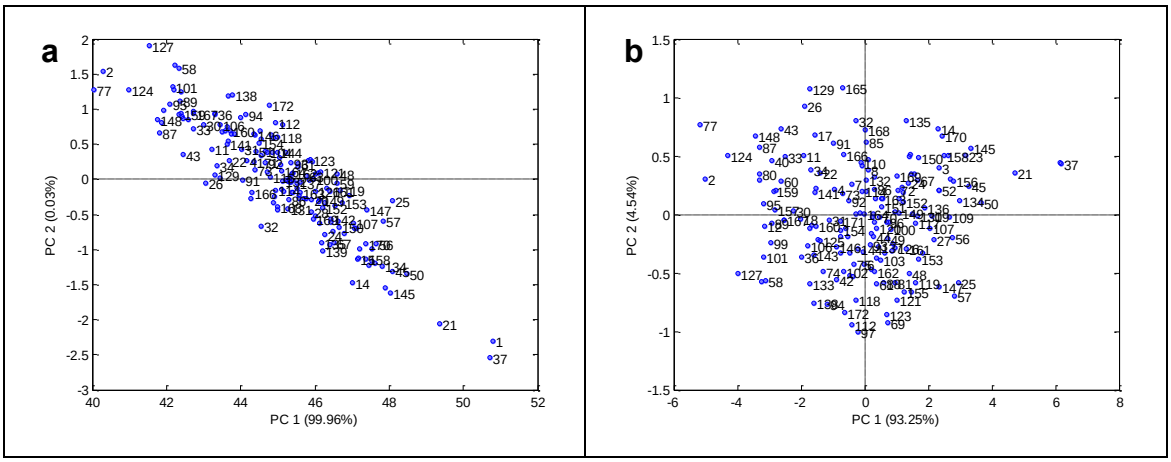

Figura 10 - Gráficos de dispersão a partir dos dois primeiros componentes principais (PC1 e PC2) para os dados de bagaço coletados no meio da safra sem nenhum pré-tratamento (a) e para os dados de bagaço coletados no meio da safra após serem centrados na média (b).

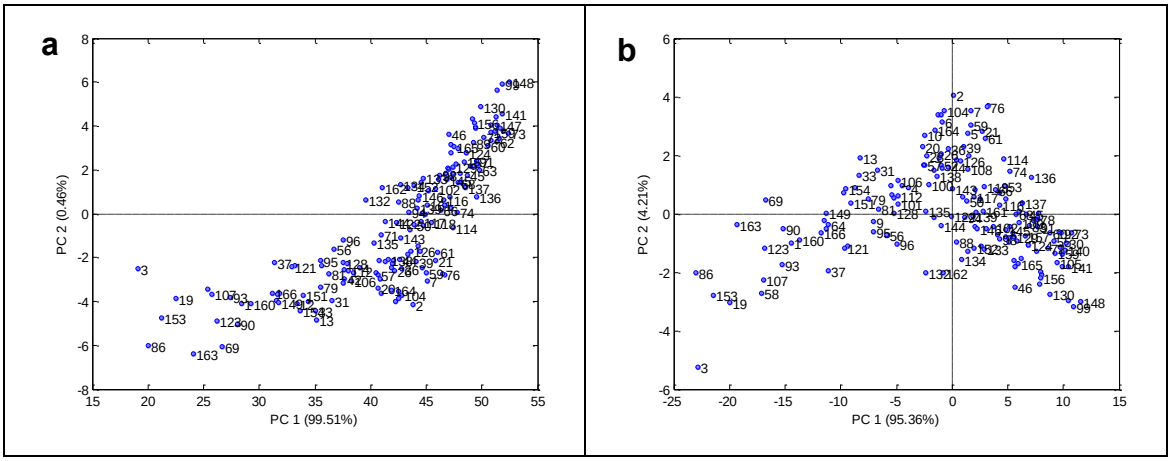

Figura 11 - Gráficos de dispersão a partir dos dois primeiros componentes principais (PC1 e PC2) para os dados de colmo sem nenhum pré-tratamento (a) e para os dados de colmo após serem centrados na média (b).

Para verificar a presença de pontos atípicos foram feitos os gráficos de Leverage versus Resíduos estudentizados (MARTENS e NAES, 1989) para cada um dos modelos escolhidos nos três bancos de dados em estudo, bagaço coletados no início e no meio da safra e colmo (Figuras 8, 9 e 10 respectivamente). Foram removidas todas as amostras anômalas com valores dos Resíduos estudentizados maiores que ±3 e/ou altos valores de Leverage. Posteriormente foi construído um novo gráfico e então foram removidas as novas amostras que apresentaram a mesma condição anterior. O processo foi repetido no máximo três vezes (FERREIRA, 2015).

Como nos dados de bagaço do início da safra dois modelos apresentaram os mesmos valores de R^2 e *RMSECV* optou-se pelo modelo com os dados centrados na média, pois ele apresentou melhor desempenho após a remoção das amostras anômalas (Tabela 3, Figura 12).

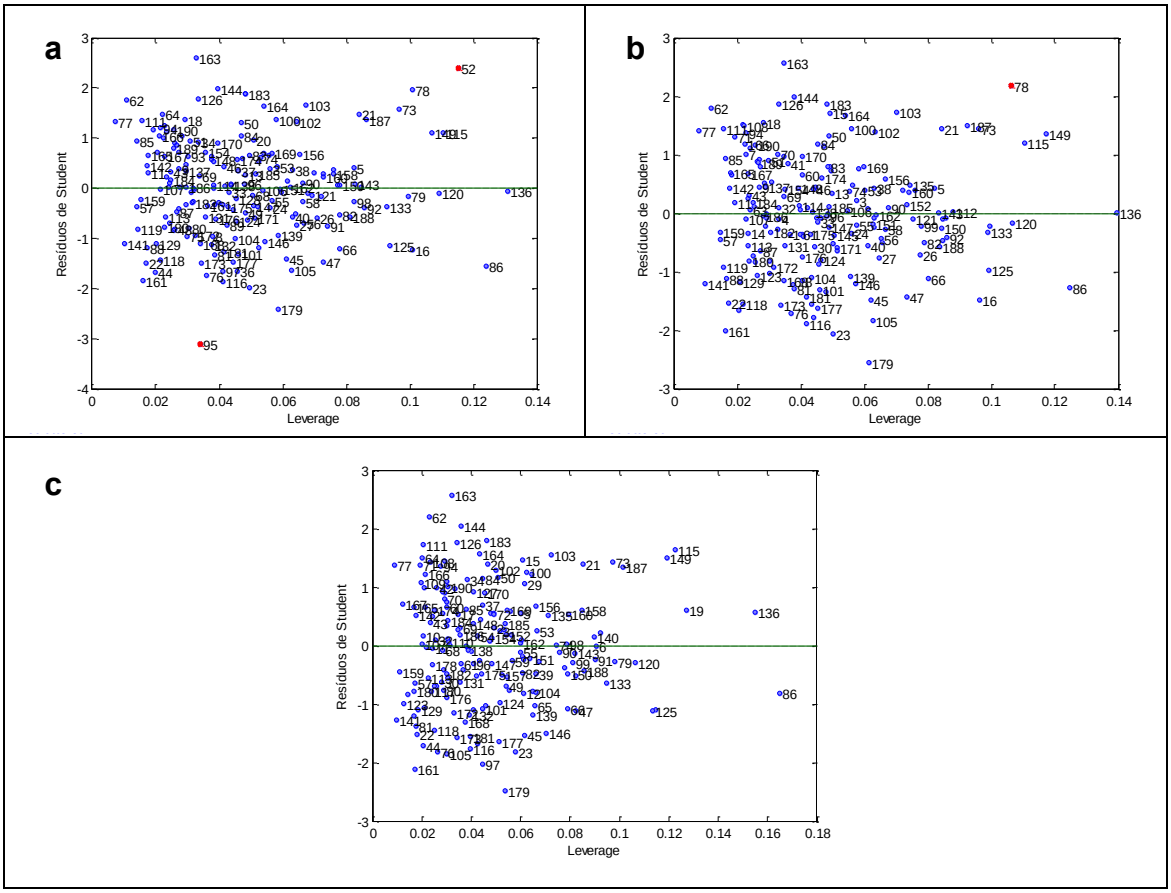

Figura 12 - Gráfico de Leverage versus Resíduos de Student para os dados de bagaço coletados no início da safra, centrados na média e com todas as amostras (a), após remover duas amostras (b) e após remover três amostras (c).

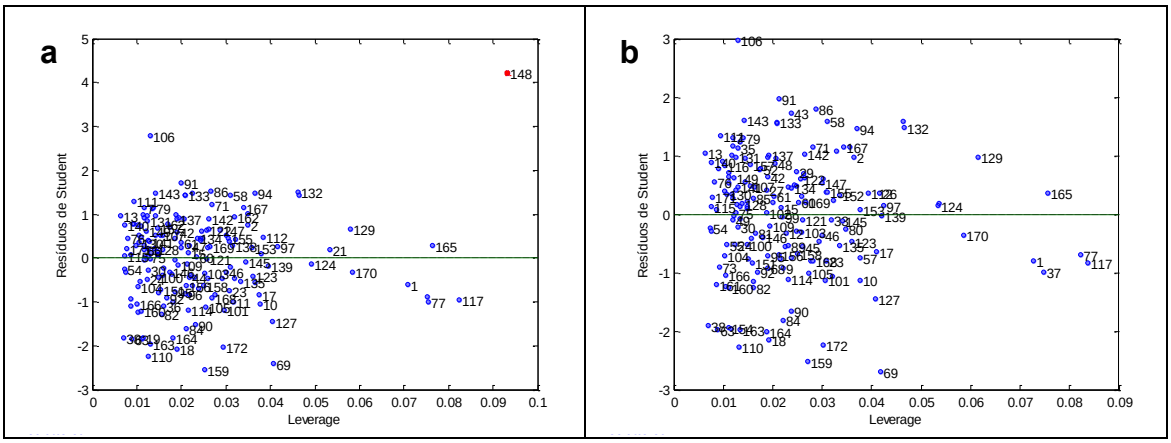

Figura 13 - Gráfico de Leverage versus Resíduos de Student para os dados de bagaço coletados no meio da safra, centrados na média e com todas as amostras (a) e após remover uma amostra.

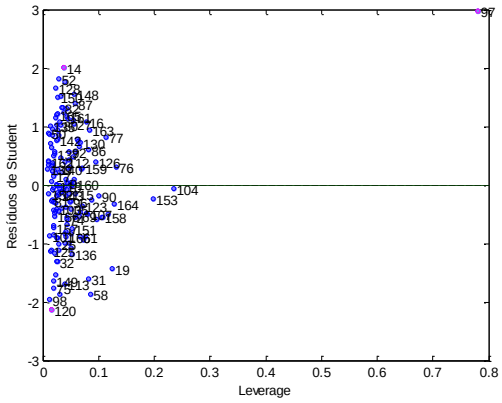

Figura 14 - Gráfico de Leverage versus Resíduos de Student para os dados de colmo com todas as amostras.

A Tabela 5 mostra os novos valores de *R²* e *RMSECV* após a detecção e retirada das amostras que estavam prejudicando o desempenho dos modelos previamente construídos.

Tabela 5 - Valores de *RMSECV* e *R²* para os melhores modelos PLS de cada um dos bancos de dados para predição do teor de fibra (FIB).

Dados	Pré-tratamento	Amostras Removidas	VL	RMSECV	R²
bagaço coletado no início da safra	centrar na média	3	9	1,69	0,35
bagaço coletado no meio da safra	-	1	4	1,84	0,41
colmo coletado no meio da safra	-	-	8	1,67	0,47

Verifica-se que o desempenho dos modelos para predizer fibra foi melhor ao utilizar dados coletados no meio da safra quando comparado com os dados coletados no início da safra. Esse resultado se deve ao fato de que no início da safra a planta ainda estava com 10 meses de idade e nesse período

ainda há muita umidade no solo, contribuindo para um maior teor de água e consequentemente um menor teor de fibra na cana-de-açúcar. Enquanto que, no meio da safra, a cana-de-açúcar passou pelo período de estresse hídrico no qual ela perde água e ocorre a estabilização do acúmulo de fibra.

Para predizer o teor de fibra é recomendado utilizar dados de colmo visto que os modelos construídos a partir desses dados apresentaram os melhores resultados, tanto de coeficiente de determinação quanto do *RMSECV*. Além disso, a preparação para a leitura do colmo é muito mais prática e rápida que a preparação para a leitura de bagaço.

A partir desses resultados, foi ajustado o modelo PLS, sem nenhum pré-tratamento e com oito variáveis latentes para o conjunto de dados que não participou do treinamento do modelo, isto é, foi realizada a validação externa. O valor do coeficiente de determinação foi igual a 0,32 e o valor do coeficiente de correlação foi igual a 0,56. O *RMSEP* apresentou valor igual a 2,7 que pode ser considerado um valor pequeno já que ele é aproximadamente quatro vezes menor que o valor mínimo do teor de fibra (Tabela 1). O gráfico de valores reais versus valores preditos é apresentado na Figura 15.

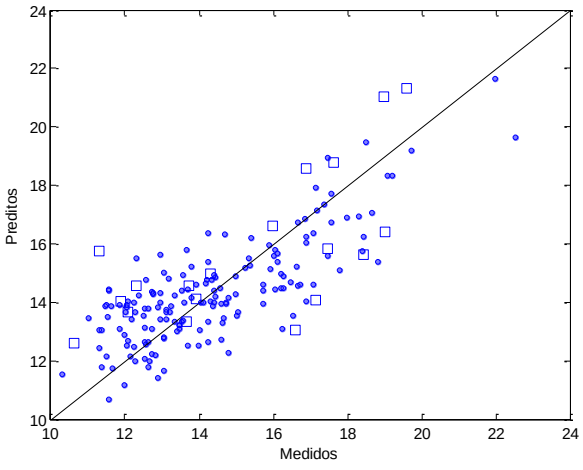

Figura 15 - Gráfico dos valores do teor de fibra real versus valores do teor de fibra predito a partir do modelo PLS em amostras de colmo sem nenhum pré-tratamento. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

Mesmo obtendo um valor do coeficiente de determinação pequeno, o modelo construído pode ser usado como triagem para ranquear os clones de acordo com o interesse dos melhoristas já que esse modelo pode ser

considerado uma boa ferramenta para identificar os maiores e os menores valores do teor de fibra, como pode ser observada na Tabela 6.

A matriz de confusão mostra que o modelo PLS escolhido retorna valores preditos capazes de classificar os clones, obtendo uma acurácia de 0,7 ou 0,8 dependendo da quantidade de amostras que se deseja selecionar, além de apresentar valores pequenos da taxa de falso positivo e com precisão variando de 0,43 a 0,8.

Tabela 6 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais do teor de fibra e a classificação obtida através dos valores preditos obtidos pelo modelo PLS construído com dados de colmo sem nenhum tratamento.

	SR=10 (50% do total)		SR=7 (fibra > 17%)		SR=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	11	2	13	2
S (Real)	2	8	4	3	2	3
Acurácia	0,80		0,70		0,80	
Taxa de falso positivo	0,20		0,29		0,40	
Precisão	0,80		0,43		0,60	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

A Tabela 7 mostra os valores dos coeficientes de correlação de diferentes modelos construídos a partir dos dados do conjunto de validação e utilizando os métodos PLS, RR-BLUP e BLASSO. Ao comparar esses valores percebe-se que o método RR-BLUP apresentou o melhor desempenho após aplicar os pré-tratamentos MSC, 1ª Derivada e Centrar na Média.

Tabela 7 - Tabela com os coeficientes de correlação para predição de fibra a partir de dados de colmo para os modelos PLS, RR-BLUP e BLASSO.

	Coeficiente de correlação		
Pré-tratamentos	PLS	RR-BLUP	BLASSO
Nenhum	**0,56**	0,43	0,45
Centrar na Média (CM)	0,54	0,53	0,62
1ª Derivada + CM	0,54	0,68	0,58
2ª Derivada + CM	0,41	0,60	0,60
MSC + CM	**0,56**	0,69	**0,67**
MSC + 1ª Derivada + CM	0,43	**0,70**	0,53
MSC + 2ª Derivada + CM	0,43	0,68	**0,67**

A Figura 16 a seguir mostra o gráfico de valores reais versus valores preditos e a Tabela 8 mostra a matriz de confusão construídos a partir do modelo RR-BLUP.

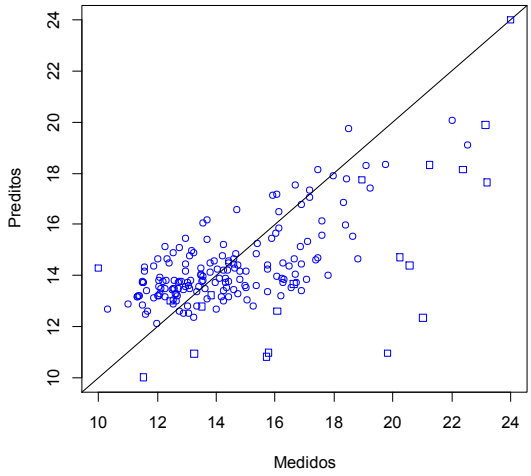

Figura 16 - Gráfico dos valores do teor de fibra real versus valores do teor de fibra predito a partir do modelo RR-BLUP em amostras de colmo após aplicar MSC e centrar na média. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

Tabela 8 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais do teor de fibra e a classificação obtida através dos valores preditos obtidos pelo modelo RR-BLUP construído com dados de colmo após aplicar MSC, 1ª Derivada e centrar na média.

	SR=10 (50% do total)		SR=7 (fibra > 17%)		SR=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	11	2	14	1
S (Real)	2	8	2	5	1	4
Acurácia	0,80		0,80		0,90	
Taxa de falso positivo	0,20		0,29		0,20	
Precisão	0,80		0,71		0,80	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Observado a matriz de confusão (Tabela 8) percebe-se que o modelo RR-BLUP apresenta acurácia e precisão maior ou igual aos obtidos a partir do modelo PLS (Tabela 6). Os valores da taxa de falso positivo obtidos nos dois modelos foram semelhantes.

4.3. Teor de sacarose aparente (PC)

Também foi avaliado o teor de sacarose aparente (PC) presente no bagaço e no colmo da cana-de-açúcar, algumas análises descritivas estão apresentadas na Figura 17 e na Tabela 9. Os valores de PC variaram entre 1,78% a 12,20% no início da safra. O percentual mínimo no meio da safra foi igual a 3,11%e o máximo igual a 16,89%. Esse aumento no teor de sacarose aparente é devido ao estágio mais avançado de maturação dos clones após passar pelo estresse hídrico causado pelos meses de seca.

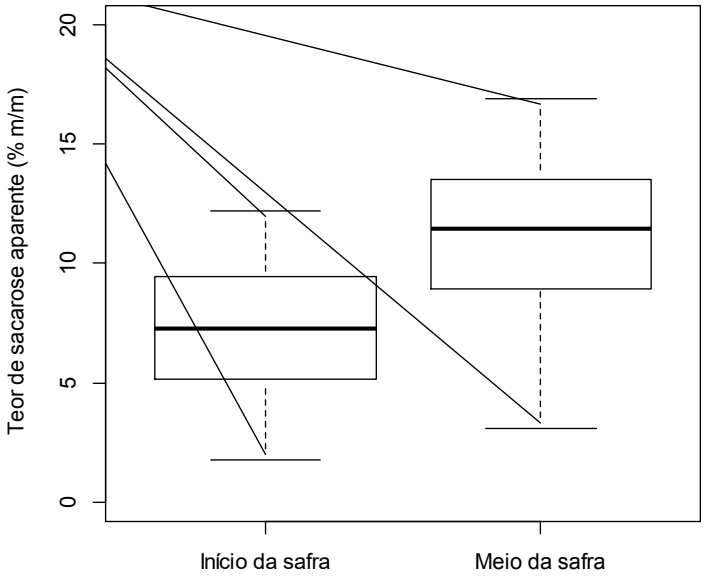

Figura 17 - Boxplots com os valores do teor de sacarose aparente (PC) obtidos a partir da análise tecnológica realizadas em amostras de bagaço no início e em amostras de bagaço no meio da safra.

Tabela 9 - Análises descritivas do caractere teor de sacarose aparente para os dados de bagaço coletados no início da safra e no meio da safra e para os dados de colmo coletados no meio da safra.

	Início da safra	Meio da safra
Média	7,24	11,27
Mediana	7,27	11,48
Mínimo	1,78	3,11
Máximo	12,20	16,89
Desvio Padrão	2,54	2,75

Verifica-se que nem todas as amostras podem ser classificadas como maduras segundo DEUBER (1988) e FERNANDES (2000), pois apresentaram teor de PC menor que 13%. Isso ocorreu, pois a população em estudo é de cana energia, portanto muitos clones apresentam baixo teor de sacarose já que são destinados a queima ou produção de álcool de segunda geração, ou seja, são materiais fibrosos e por isso sem aptidão para o acumulo de sacarose.

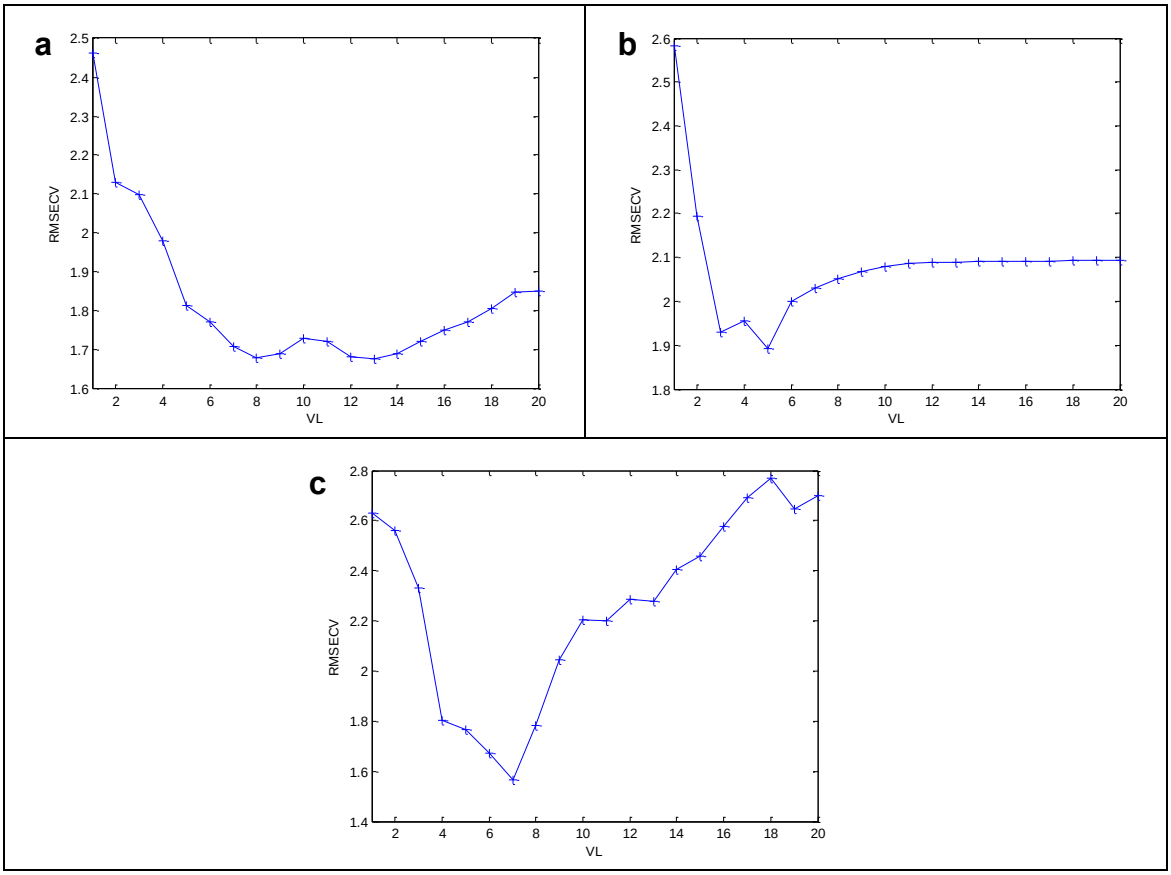

Figura 18 - Número de variáveis latentes (VL) versus a raiz do erro quadrático médio de validação cruzada (*RMSECV*) em dados de bagaço coletados no início da safra após aplicar MSC e centrar na média (a), em dados de bagaço coletados no meio da safra após aplicar MSC e centrar na média (b) e em dados de colmo após aplicar MSC, primeira derivada e centrar na média (c).

Após analisar a Figura 18 foi escolhida a melhor quantidade de variáveis latentes para a construção dos modelos. Para os dados de bagaço coletado no início da safra foi utilizada VL igual a oito, para os dados de bagaço coletado no meio da safra foi utilizada VL igual a cinco e finalmente para os dados de colmo o número de VL escolhido foi sete.

Tabela 10 - Valores de *RMSECV* e de *R²* para diferentes tratamentos em dados de bagaço coletados no início da safra, bagaço coletados no meio da safra e colmo da cana-de-açúcar para predição do PC utilizando o modelo PLS.

	bagaço do início da safra			bagaço do meio safra			colmo do meio da safra		
Pré-tratamentos	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**
Nenhum	8	1,72	0,54	5	2,01	0,47	7	3,28	0,20
Centrar na Média (CM)	8	1,77	0,52	5	2,05	0,44	7	2,94	0,27
1ª Derivada + CM	8	1,79	0,52	5	2,55	0,19	7	1,83	0,58
2ª Derivada + CM	8	2,28	0,29	5	3,05	0,04	7	2,40	0,35
MSC + CM	**8**	**1,68**	**0,57**	**5**	**1,89**	**0,52**	7	1,73	0,61
MSC + 1ª Derivada + CM	8	1,81	0,52	5	3,05	0,05	**7**	**1,56**	**0,68**
MSC + 2ª Derivada + CM	8	2,27	0,30	5	2,51	0,22	7	1,95	0,52

Ao modelar os dados utilizando o método PCR (Tabela 11) verifica-se um comportamento semelhante ao ocorrido usando o PLS (Tabela 10), no entanto, os coeficientes de determinação calculados são menores. Dessa forma, o modelo PLS se mostra superior ao PCR e por isso ele é o mais utilizado para realizar calibração multivariada em dados de NIR (FERREIRA, 2015).

Tabela 11 - Valores de *RMSECV* e de *R²* para diferentes tratamentos em dados de bagaço coletados no início da safra, bagaço coletados no meio da safra e colmo da cana-de-açúcar para predição do PC utilizando o modelo PCR.

	bagaço do início da safra			bagaço do meio safra			colmo do meio da safra		
Pré-tratamentos	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**
Nenhum	8	2,04	0,36	5	2,01	0,47	7	3,24	0,03
Centrar na Média (CM)	8	1,99	0,39	5	2,04	0,44	7	3,10	0,03
1ª Derivada + CM	8	1,77	0,52	5	2,55	0,19	7	2,46	0,26
2ª Derivada + CM	8	2,49	0,05	5	3,05	0,04	7	2,75	0,04
MSC + CM	8	1,97	0,40	**5**	**1,89**	**0,52**	7	2,33	0,29
MSC + 1ª Derivada + CM	**8**	**1,74**	**0,53**	5	2,50	0,21	**7**	**1,82**	**0,56**
MSC + 2ª Derivada + CM	8	2,50	0,05	5	3,05	0,05	7	2,67	0,07

Para verificar a presença de pontos atípicos nos dados foram construídos gráficos de Leverage versus Resíduos de Student (MARTENS e NAES, 1989) após a escolha do melhor modelo para cada um dos conjuntos de dados. Os gráficos são apresentados a seguir (Figuras 18, 19 e 20).

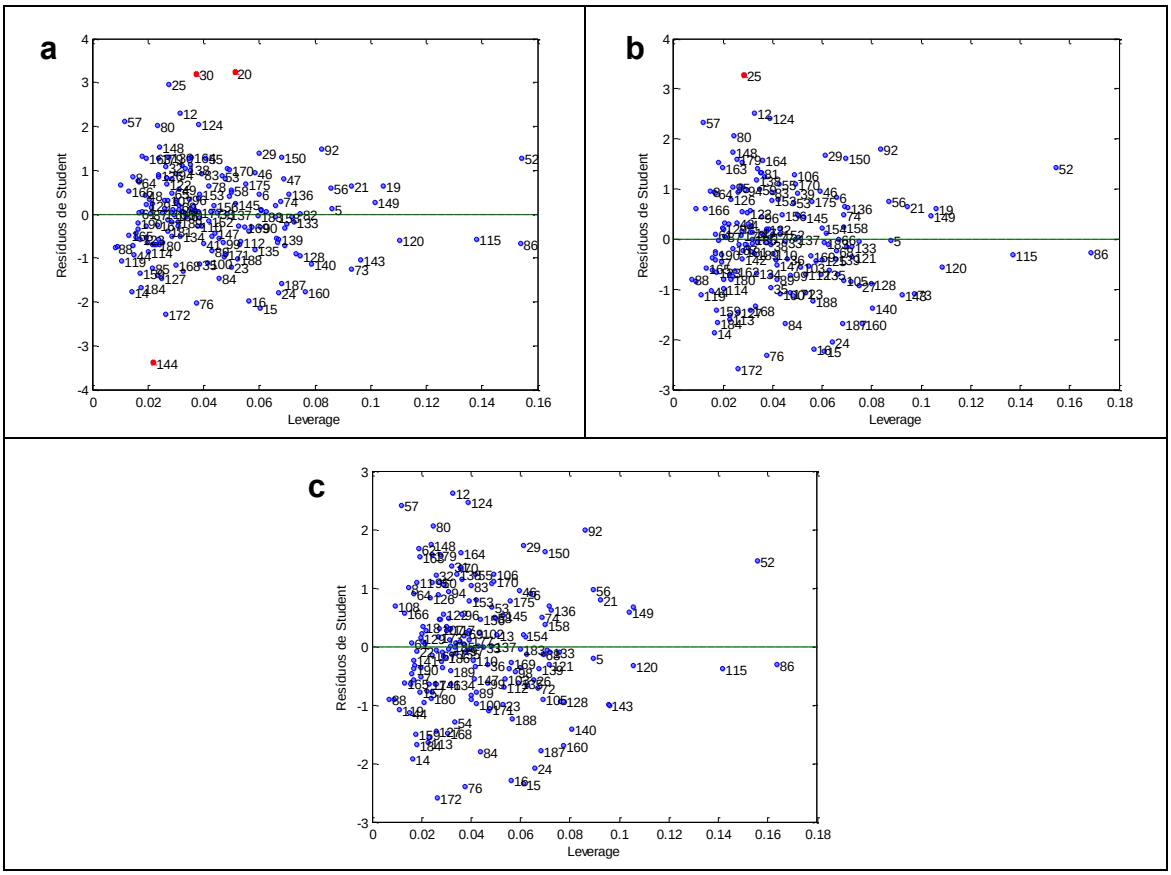

Figura 19 - Gráfico de Leverage versus Resíduos de Student para os dados de bagaço congelado coletado no início da safra com todas as amostras (a), após remover três amostras (b) e após remover quatro amostras (c).

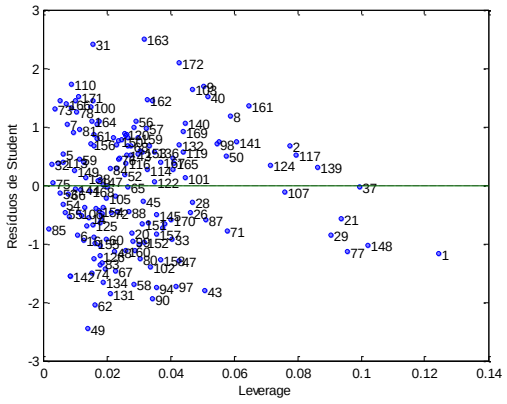

Figura 20 - Gráfico de Leverage versus Resíduos de Student para os dados de bagaço coletado no meio da safra com todas as amostras.

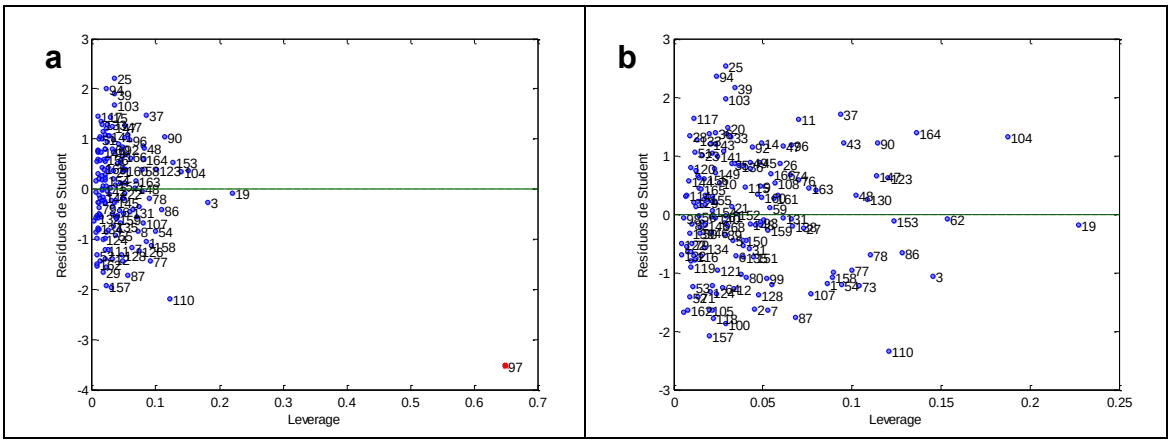

Figura 21 - Gráfico de Leverage versus Resíduos de Student para os dados de colmo com todas as amostras (a) e após remover uma amostra (b).

A Tabela 12 mostra os melhores modelos para cada um dos conjuntos de dados. Com a remoção das amostras anômalas identificadas a partir do gráfico de Leverage versus Resíduos de Student os modelos melhoram o desempenho visto que os valores dos coeficientes de determinação (R^2) aumentaram e os erros (*RMSECV*) diminuíram.

Tabela 12 - Valores de *RMSECV* e *R²* para os melhores modelos de cada um dos bancos de dados para predição do teor de sacarose aparente (PC).

Dados	Pré-tratamento	Amostras Removidas	VL	RMSECV	R^2
bagaço coletado no início da safra	MSC + CM	4	8	1,47	0,66
bagaço coletado no meio da safra	MSC + CM	0	5	1,89	0,52
colmo coletado no meio da safra	MSC + 1ªD + CM	1	7	1,48	0,71

Ao utilizar dados de bagaço coletados no início da safra foi possível obter um melhor desempenho do modelo quando comparado com os dados coletados no meio da safra.

Para predizer o teor de sacarose aparente é aconselhável utilizar dados de colmo coletados no início da safra visto que os modelos construídos a partir desses dados apresentaram resultados superiores aos resultados obtidos a partir do bagaço. Além disso, a preparação para a leitura dos dados de colmo é muito mais prática e rápida.

Assim, foi ajustado o modelo PLS após aplicar MSC, primeira derivada e centragem na média utilizando sete variáveis latentes para o conjunto de dados colmo que não participou do treinamento do modelo, isto é, foi realizada a validação externa. O valor de R^2 encontrado foi igual a 0,64 e o valor do *RMSEP* foi igual a 3,07 que é um pouco maior que o valor mínimo de PC. O

Gráfico de valores medidos versus valores preditos é apresentado na Figura 22.

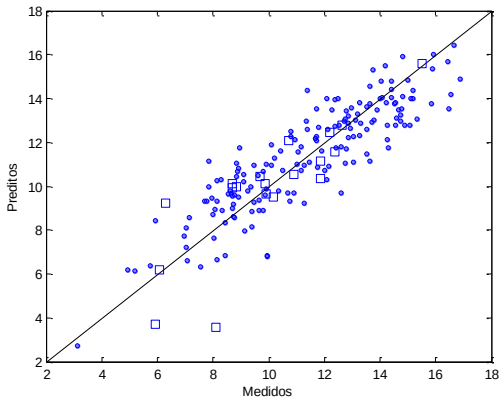

Figura 22 - Gráfico dos valores do teor de sacarose aparente real versus valores do teor de sacarose aparente preditos via NIR a partir da construção do modelo PLS em amostras de colmo após aplicar MSC, 1ª Derivada e centrar na média. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

O modelo apresentado mostra-se como uma boa ferramenta para ranquear os clones (Tabela 13) mesmo apresentando um valor do coeficiente de determinação não muito alto (0,65).

A matriz de confusão mostra que o modelo escolhido retorna valores preditos capazes de classificar os clones já que apresentou uma acurácia alta variando entre 0,8 e 0,9, dependendo do número de clones selecionados, além de baixos valores da taxa de falso positivo e alta precisão (Tabela 13).

Tabela 13 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores medidos do teor de sacarose aparente e a classificação obtida a partir dos valores preditos do teor de sacarose aparente pelo modelo PLS construído com dados de colmo após aplicar MSC, 1ª Derivada e centrar os dados na média.

	SR=10 (50% do total)		SR=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	14	1
S (Real)	2	8	1	4
Acurácia	0,80		0,90	
Taxa de falso positivo	0,20		0,20	
Precisão	0,80		0,80	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Também foram construídos os modelos RR-BLUP e BLASSO com os dados de colmo após aplicar diferentes pré-tratamentos para comparar com modelo PLS escolhido anteriormente. Os valores dos coeficientes de correlação estão apresentados na Tabela 14, na qual mostra que ao utilizar o modelo BLASSO após aplicar MSC e centrar os dados na média é possível obter um modelo eficiente. O gráfico de valores medidos versus valores preditos e a matriz de confusão para o modelo BLASSO são apresentados na Figura 23 e na Tabela 15 abaixo.

Tabela 14 - Tabela com os coeficientes de correlação para predição do teor de sacarose aparente a partir de dados de colmo para os modelos PLS, RR-BLUP e BLASSO.

	Coeficiente de correlação		
Pré-tratamentos	PLS	RR-BLUP	BLASSO
Nenhum	0,63	0,59	0,58
Centrar na Média (CM)	0,55	0,56	0,56
1ª Derivada + CM	**0,83**	0,70	0,67
2ª Derivada + CM	0,74	0,54	0,49
MSC + CM	0,80	0,77	**0,83**
MSC + 1ª Derivada + CM	0,80	**0,80**	0,75
MSC + 2ª Derivada + CM	0,77	0,57	0,50

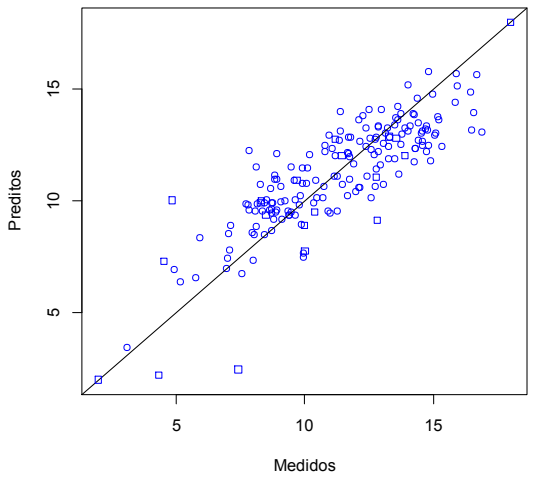

Figura 23 - Gráfico dos valores do teor de sacarose aparente real versus valores do teor de sacarose aparente preditos via NIR a partir do modelo BLASSO construído em amostras de colmo após aplicar MSC e centrar os dados na média. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

A Tabela 14 mostra que a combinação dos pré-tratamentos 1ª Derivada e Centrar média utilizando o método PLS resulta em um mesmo valor do

coeficiente de correlação (0,83) comparado com o valor encontrado após aplicar o método BLASSO e aplicar os pré-tratamentos MSC, 1ª Derivada e Centrar na média.

Tabela 15 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais do teor de sacarose aparente e a classificação obtida a partir dos valores preditos do teor de sacarose aparente pelo modelo BLASSO construído com dados de colmo após aplicar 1ª Derivada e centrar os dados na média.

Classificação	SR=10 (50% do total)		SR=5 (25% do total)	
	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	7	3	14	1
S (Real)	3	7	1	4
Acurácia	0,70		0,90	
Taxa de falso positivo	0,30		0,20	
Precisão	0,70		0,80	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Comparando os resultados das matrizes de confusão obtidos a partir do modelo PLS e a partir do modelo BLASSO percebe-se que o modelo BLASSO apresentou valores de acurácia e precisão um pouco menor, além de um valor da taxa de falso positivo um pouco maior ao selecionar 50% do total de amostras. E ao selecionar 25% do total de amostras o desempenho dos dois modelos foi o mesmo.

4.4. Teor de lignina (LIG)

Um resumo dos dados de lignina é apresentado na Figura 23 e na Tabela 16 que mostra as principais análises descritivas. Para a construção dos modelos de predição do teor de lignina (LIG) na fibra, foram utilizados dados de bagaço obtidos em leituras de amostras de bagaço seco, moído e peneirado. Inicialmente foi verificado o melhor valor do número de varáveis latentes (Figura 25). Em seguida, foram construídos diferentes modelos alternando os tipos de pré-tratamentos e tipos de métodos, PLS e PCR (Tabela 17).

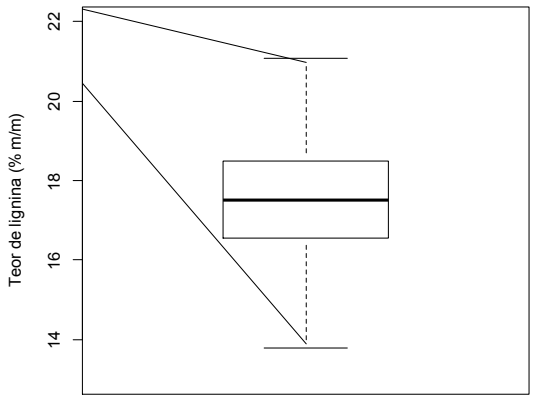

Figura 24 - Boxplot com os valores do teor de lignina (LIG) obtidos a partir da predição via NIR realizada pela celignis em dados de bagaço seco coletado no meio da safra.

Tabela 16 - Análises descritivas do caractere teor de lignina obtidos a partir da predição via NIR realizada pela celignis em dados de bagaço seco coletado no meio da safra.

	Bagaço seco
Média	17,48
Mediana	17,51
Mínimo	13,79
Máximo	21,08
Desvio Padrão	1,43

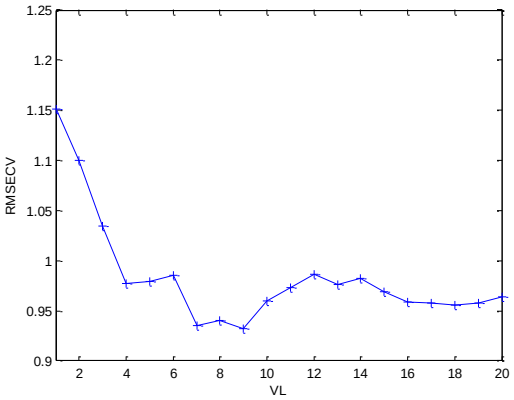

Figura 25 - Número de variáveis latentes versus *RMSECV* em dados de bagaço seco após aplicar 1ªDerivada e centrar na média.

Tabela 17 - Valores de *RMSECV* e de *R²* para diferentes tratamentos em dados de bagaço seco para predição de lignina utilizando os modelos PLS e PCR.

	PLS			PCR		
Pré-tratamentos	**VL**	**RMSECV**	**R²**	**VL**	**RMSECV**	**R²**
Nenhum	9	1,20	0,33	9	1,19	0,35
Centrar na Média (CM)	9	1,10	0,41	9	1,05	0,44
1ª Derivada + CM	**9**	**0,93**	**0,57**	9	1,00	0,48
2ª Derivada + CM	9	1,09	0,43	9	1,20	0,26
MSC + CM	9	0,96	0,54	9	1,02	0,46
MSC + 1ª Derivada + CM	9	0,99	0,52	**9**	**1,00**	**0,49**
MSC + 2ª Derivada + CM	9	1,13	0,40	9	1,22	0,24

Pelos resultados observados na Tabela 17 verifica-se que o melhor modelo para predição de lignina é utilizando o PLS após aplicar a 1ª Derivada e centrar os dados na média. Em outro trabalho no qual também foram testados diferentes pré-processamentos a melhor combinação de pré-tratamentos foi 2ª derivada e centrar na média (ASSIS, 2010).

Foi construído o gráfico de Leverage versus Resíduos de Student para poder identificar e remover as amostras anômalas que podem prejudicar o desempenho do modelo escolhido (MARTENS e NAES, 1989). A Figura 26 mostra que nenhum amostra anômala foi entrada.

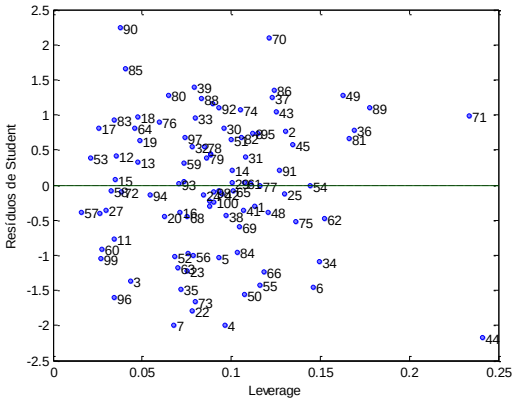

Figura 26 - Gráficos de Leverage versus Resíduos de Student para os dados de bagaço seco com todas as amostras.

Após realizar a validação externa foi obtido um coeficiente de determinação (R²) igual a 0,70, coeficiente de correlação igual a 0,84 e RMSEP igual a 0,88 que é um valor de erro, aproximadamente, 16 vezes menor que o

valor mínimo do teor de lignina (Tabela 16), indicando um bom modelo para predição. O gráfico com os valores reais versus valores preditos é apresentado a seguir (Figura 27).

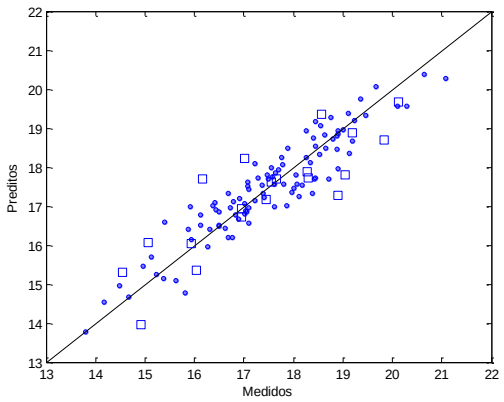

Figura 27 - Gráfico dos valores do teor de lignina real versus valores do teor de lignina preditos via NIR a partir do modelo PLS ajustado em amostras de bagaço seco. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

Utilizando o modelo escolhido é possível classificar os clones de forma que os resultados serão bem próximos da classificação real, pois, pela matriz de confusão (Tabela 18) observa-se uma medida de acurácia igual a 0,80 no caso de selecionar 50% ou 25% da amostra. Além disso, a medida de false positivo rate apresentou um valor pequeno ao selecionar 50% da amostra e a precisão variou de 0,60 a 0,80 dependendo da quantidade de clones selecionados.

Tabela 18 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais de lignina e a classificação obtida através dos valores preditos obtidos pelo modelo PLS construído com dados de bagaço seco coletado no meio da safra.

	S=10 (50% do total)		S=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	13	2
S (Real)	2	8	2	3
Acurácia	0,80		0,80	
False positive rate	0,20		0,40	
Precisão	0,80		0,60	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Também foram construídos os modelos RR-BLUP e BLASSO a fim de predizer o teor de lignina a partir dos dados de bagaço seco. Esses modelos apresentaram um bom desempenho principalmente após aplicar os pré-tratamentos: MSC, 1ª Derivada e centrar na média (Tabela 20), superando os modelos construídos a partir do método PLS. O gráfico de valores medidos versos valores preditos (Figura 28) e as matrizes de confusão (Tabelas 20 e 21) dos modelos RR-BLUP e BLASSO estão apresentados a seguir.

Tabela 19 - Tabela com os coeficientes de correlação para predição do teor de lignina a partir de dados de bagaço seco para os modelos PLS, RR-BLUP e BLASSO.

	Coeficiente de correlação		
Pré-tratamentos	PLS	RR-BLUP	BLASSO
Nenhum	0,86	-0,10	0,12
Centrar na Média (CM)	**0,88**	0,79	0,84
1ª Derivada + CM	0,84	0,88	0,88
2ª Derivada + CM	0,83	0,89	0,89
MSC + CM	0,83	0,86	0,88
MSC + 1ª Derivada + CM	0,85	**0,91**	**0,91**
MSC + 2ª Derivada + CM	**0,88**	0,86	0,86

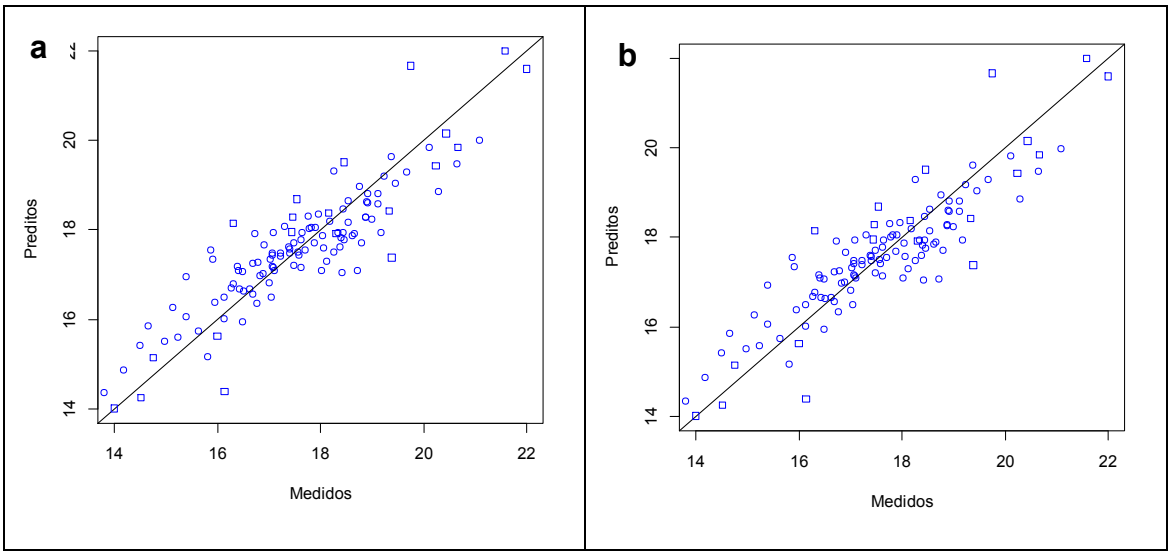

Figura 28 - Gráfico dos valores do teor de lignina versus valores do teor de lignina preditos via NIR a partir do modelo RR-BLUP (a) e do modelo BLASSO (b) construídos em amostras de bagaço seco após aplicar MSC, 1ª Derivada e Centrargem na média. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

Tabela 20 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais do teor de lignina e a classificação obtida a partir dos valores preditos do teor de lignina pelo modelo RR-BLUP construído com dados de bagaço seco após aplicar MSC, 1ª Derivada e Centragem na média.

	SR=10 (50% do total)		SR=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	14	1
S (Real)	2	8	1	4
Acurácia	0,80		0,90	
Taxa de falso positivo	0,20		0,20	
Precisão	0,80		0,80	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Tabela 21 - Matriz de confusão entre a classificação verdadeira feita a partir dos valores reais do teor de lignina e a classificação obtida a partir dos valores preditos do teor de lignina pelo modelo BLASSO construído com dados de bagaço seco após aplicar MSC, 1ª Derivada e Centragem na média.

	SR=10 (50% do total)		SR=5 (25% do total)	
Classificação	NS (Preditos)	S (Preditos)	NS (Preditos)	S (Preditos)
NS (Real)	8	2	14	1
S (Real)	2	8	1	4
Acurácia	0,80		0,90	
Taxa de falso positivo	0,20		0,20	
Precisão	0,80		0,80	

*SR=Clones selecionados a partir do valor real, S=Clones selecionados e NS=Clones não selecionados.

Ao comparar os modelos RR-BLUP e BLASSO percebe-se que não houve diferença entre as medidas do coeficiente de correlação e também entre as medidas de acurácia, taxa de falso positivo e precisão. Logo, para predizer lignina a partir de dados de bagaço seco é possível escolher entre esses dois métodos.

5. CONCLUSÕES

Para predizer o teor de fibra (FIB) e o teor de sacarose aparente (PC) é aconselhável utilizar dados de colmo devido, principalmente, ao maior poder preditivo. Foi possível construir modelos eficientes para predição do teor de lignina a partir dos dados fenotípicos obtidos via empresa celignis.

O modelo construído a partir do método RR-BLUP após aplicar os pré-tratamentos MSC, 1ª Derivada e Centrar na média foi o melhor modelo para predizer o teor de fibra. Para predizer o teor de sacarose aparente é recomendo utilizar modelos construídos utilizando o método PLS ou o método BLASSO. Para predizer lignina os melhores modelos foram aqueles utilizando os métodos RR-BLUP e BLASSO após os dados serem submetidos aos pré-tratamentos MSC, 1ª Derivada e Centrar na média.

Todos os modelos escolhidos para predizer cada um dos caracteres em estudo apresentaram alto valor da medida de acurácia e da medida de precisão, além de valores pequenos da medida da taxa de falso positivo.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ASSIS, C., Previsão do teor de lignina em cana-de-açúcar usando espectroscopia no infravermelho próximo e métodos quimiométricos. 2010. 87 f. Dissertação (Mestrado em Agroquímica), Universidade Federal de Viçosa,2010.

BARTHUS, R. C.; MAZO, L. H.; POPPI. R. J. Determinação simultânea de NADH e ácido ascórbico usando voltametria de onda quadrada com eletrodo de carbono vítreo e calibração multivariada. Eclética química30, no. 4, p. 51-58, 2005.

BARLOW, J. L.; BOSNER, N.; DRMAC, Z. A new stable bidiagonal reduction algorithm. Linear Algebra and Its Applications, p. 35-84, 2005.

BLANCO, M.; VILLARROYA, I. NIR spectroscopy: a rapid-response analytical tool. Trends in analytical chemistry. v. 21, n. 4, p. 240-250, 2002.

BRO, R. Multivariate calibration: What is in chemometrics for the analytical chemist? Analytica Chimica Acta, 500(1), p.185-194, 2003.

CANILHA, L., et al. Sacarificação da biomassa lignocelulósica através de pré-hidrólise ácida seguida por hidrólise enzimática: uma estratégia de "desconstrução" da fibra vegetal. Revista Analytica 44, p. 48-54, 2009.

CONAB: Companhia Nacional de Abastecimento. Central de informações agropecuárias: safras – cana. 2015/16. Disponível em: <http://www.conab.gov.br> acessado em 17 de janeiro, 2016.

CORDEIRO, G. M. e LIMA NETO, E. A. Modelos Paramétricos. Recife: Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, 2006.

DE GORTER, H.; JUST, D. R. The Social Costsand Benefits of Biofuels: The Intersection of Environmental, Energy and Agricultural Policy. Applied Economic Perspectives and Policy, (32), p. 4-32, 2010.

DE MUÑIZ, G. I. B. et al. Fundamentos e estado da arte da espectroscopia no infravermelho próximo no setor de base florestal. Ciência Florestal, v. 22, n. 4, p. 865-875, 2012.

DEUBER, R. Maturação da cana-de-açúcar na região sudeste do Brasil. In: SEMINÁRIO DE TECNOLOGIA DA COPERSUCAR, 1988. Piracicaba. Anais. Piracicaba: Copersucar, p. 33-40, 1988.

DONG, X.; SUN, X. Rapid determination of reducing sugar content by near infrared spectroscopy. Optoelectronic Devices and Integration. Optical Society of America, 2014.

FERNANDES, A. C. Cálculos na Agroindústria da cana de açúcar. Piracicaba, STAB: Açúcar, Álcool e Subprodutos, 2000.

FERREIRA, M. M. C. Multivariate QSAR. J. Braz. Chem. Soc., São Paulo, v.13, n.6, p.742-753, 2002.

FERREIRA, M. M. C. Qumiometria - Conceitos, Métodos e Aplicações. Campinas, SP: Editora da Unicamp, 2015.

FEUDALE, R. N. et al. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems, v 64, p. 181-92, 2002.

GONTIJO NETO, M. M., et al. Predição de proteína, matéria seca e fósforo em grãos de milho pela espectroscopia de reflectância no infravermelho próximo. Embrapa Milho e Sorgo. Boletim de Pesquisa e Desenvolvimento, 2009.

HARALD, M. and TORMOD, N. Multivariate Calibration. Wiley, Chichester, 1989.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction.(Springer Series in Statistics, Ed. 2), 2nd ed. Springer-Verlag, Stanford, CA, 2009

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. Technometrics 11, p.137-148, 1969.

KHESHGI, H. S.; PRINCE, R. C.; MARLAND, G. The Potential of Biomass Fuels in the contexto of global climate change: Focus on Transportation Fuels. Annu. Rev. Energy Environ. 2000, 25, 1999.

LOUREIRO, M.E.; BARBOSA, M.H.P.; LOPES, F.J.F.; SILVÉRIO, F.O. Sugarcane Breeding and Selection for more Efficient Biomass Conversion in Cellulosic Ethanol. In: BUCKERIDGE, M.S.; GOLDMAN, G.H. Routes to Cellulosic Ethanol. New York, Springer, cap.13, p.199-239, 2011.

MANNE, R. Analysis of 2 partial-least-squares algorithms for multivariate calibration. Chemometrics and Intelligent Laboratory Systems, p. 187-197, 2001.

MAPA: Ministério da Agricultura, Pecuária e Abastecimento. Vegetal. Cana-de-açúcar. Disponível em: <http://www.agricultura.gov.br/> acessado em 20 de janeiro, 2016.

MARTENS, H.; NAES, T. Multivariate Calibration, John Wiley & Sons, New York, 1989.

MATSUOKA, S.et al. Bioenergia da Cana. In: SANTOS, F. et al., Ed. Cana-de-açúcar: Bioenergia, Açúcar e Álcool – Tecnologias e Perspectivas. Viçosa, p. 487, 2010.

MEUWISSEN, T. H.; HAYES, B.J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157:1819–1829, 2001.

MINGOTTI, S. A. Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada. Editora UFMG, 2007.

MORGANO, M. A., et al. Determinação de umidade em café cru usando espectroscopia NIR e regressão multivariada. Ciência Tecnologia Alimentos, v. 28, n. 1, p. 12-17, 2008.

NAES, T., et al. A user friend lyguide to multivariate calibration and classification. NIR publications, 2002.

NAWI, N. M., et al. Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared.biosystems engineering 115.2, 2013.

OZAKI, Y.; MCCLURE, W.F.; CHRISTY, A. A. Near-infrared spectroscopy in food scince and technology.Wiley Inter-science, p. 408, 2007.

PARK, T.;CASELLA, G.The Bayesian LASSO. J. Am. Stat. Assoc. 103: 681–686, 2008.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspectsand Analytical Applications. J. Braz. Chem. Soc., São Paulo, v.14, n.2, 2003.

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

RESENDE, M. D. V., et al. Métodos estatísticos na seleção genômica ampla, 1st edn. Embrapa Florestas, Colombo, 2011.

ROCHA G. J. M, MARTIN C., SILVA V. F. N, GOMEZ, E. O. GONÇALVES, A. R. Mass balance of pilot-scale pretreatment of sugarcane bagasse by steam explosion followed by alkaline delignification.Bioresour. Technol. 111: 447–452, 2012.

RODRIGUES, J. A. R. Do engenho à biorrefinaria. A usina de acúcar como empreendimento industrial para a geração de produtos bioquímicos e biocombustíveis. Química Nova, Campinas, v. 34, n. 7, p. 1242-1254, 2011.

SANTCHURM, Dk, et al. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica,185.3, p. 543-558, 2012.

SANTOS, G. A., PEREIRA, A. B., KORNDORFER, G. H. Uso do sistema de análises por infravermelho próximo (NIR) para análises de matéria orgânica e fração argila em solos e teores foliares de silício e nitrogênio em cana-de-açúcar. Bioscience Journal, 26, 2010.

SANTOS, F. A., et al. Potencial da palha de cana-de-açúcar para produção de etanol. Química Nova35.5, p. 1004-1010, 2012.

SILVA, D. M. et al. Obtenção de derivado de celulose a partir do bagaço de cana-de-açúcar com potencial aplicação nas indústrias farmacêutica e cosmética. Revista de Ciências Farmacêuticas Básica e Aplicada, Araraquara, v. 32, n. 1, p. 41-45, 2011.

SILVEIRA, L. C. I., BRASILEIRO, B.P.; KIST, V.; DAROS, E.; PETERNELLI, L.A.; BARBOSA, M.H.P. Selection strategy in families of energy cane based on biomass production and quality traits. Euphytica, v. 201, p. 443-455, 2015.

SOUZA, A. P., LEITE, D. C. C, Pattathil S, Hahn MG, Buckeridge MS Composition and structure of sugarcane cell wall polysaccharides: implications for second-generation bioethanol production. BioEnergy Research, 6: 64–579, 2013.

TEÓFILO, R. F. Chemometric methods in the electrochemical studies of phenols on boron-doped diamond films. Universidade Estadual de Campinas, Campinas, 2007.

TEÓFILO, R. F. Métodos Quimiométricos: Uma Visão Geral- Conceitos básicos de quimiometria. Universidade Federal de Viçosa, Viçosa, Vol. 1, 2013.

TEW, T.L.; COBILL, R.M. Genetic Improvement of Sugarcane (Saccharum spp.) as na Energy Crop. In: VERMERRIS, W. Genetic Improvement of Bioenergy Crops. Springer, New York, cap. 9, p. 273-294, 2008.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. P. Validation of Multivariate Calibration Models in the Determination of Sugar Cane Quality Parameters by Near Infrared Spectroscopy. J. Braz. Chem. Soc. v. 18, n. 2, 2007.

WHITTAKER J.C., Thompson R, Denham MC. Marker-assisted selection using ridge regression. Genet. p. 249 – 252, 2000.

YE, W., et al. Application of near-infrared reflectance spectroscopy for determination of nutrient contents in liquid and solid manures. Transactions of the ASAE 48.5, p. 1911, 2005.

ZHENG, Y, et al. Enzymatic saccharification of dilute acid pretreated saline crops for fermentable sugar production. Applied Energy 86.11, p. 2459-2465, 2009.